ACT Research Report Series 87-14

The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence

Terry Ackerman

September 1987

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

THE ROBUSTNESS OF LOGIST AND BILOG IRT ESTIMATION PROGRAMS TO VIOLATIONS OF LOCAL INDEPENDENCE

Terry A. Ackerman

ABSTRACT

One of the important underlying assumptions of all item response theory models is that of local independence. This assumption requires that the response to an item on a test not be influenced by the response to any other items. This assumption is often taken for granted, with little or no scrutiny of individual test items for possible violations of local independence.

Ackerman and Spray (1986) proposed a dependency model with which the interaction of such factors as the amount and direction of item dependency, item difficulty and discrimination, and item order or sequence effects could be simulated. In this study item response data were generated with varying degrees of response dependency using their model. These data were used to determine the robustness of the IRT calibration programs LOGIST and BILOG to violations of local independence. Results suggest that calibrated dependent item discrimination parameters tend to be overestimated, and difficulty estimates tend to become homogeneous. Ability estimates, however, were affected as the degree of dependency increased.

The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence

Although it is one of the basic assumptions underlying item response theory (IRT), local independence of item responses is often taken for granted, with little attention paid to determine if the process of responding to one item influences the response(s) to other item(s). Yet violations of this assumption can easily occur when several items are embedded in the same passage, or when items contain multiple parts. Local independence is violated whenever the response process of one item provides the necessary cognitive schema to trigger a response to a subsequent item.

The purpose of this paper is to examine the robustness of the IRT calibration programs BILOG (Mislevy & Bock, 1984) and LOGIST (Wingersky, Barton, & Lord, 1982) to varying degrees of local dependence. Dependency was imposed upon a set of real test data using the model developed by Ackerman and Spray (1986).

Model Definition

Ackerman and Spray (1986) proposed an item dependency model which is based upon a finite, two state (0 or 1, incorrect or correct) Markov process. In the model $P_j(\theta_i)$ is defined as the probability an examinee with ability, θ_i, will answer item j correctly, independently of any other test item. $P_j(\theta_i)$ can be determined using any response function, however for the purposes of this study, it was defined by the two parameter logistic IRT model.

The dependency model is defined by a transition matrix between any two items, j - 1 and j, in a k-item test:

		jth item 0	jth item 1
jth - 1 item	0	$1 - \alpha P_j(\theta_i)$	$\alpha P_j(\theta_i)$
	1	$\beta Q_j(\theta_i)$	$1 - \beta Q_j(\theta_i)$

In this model, $\alpha P_j(\theta_i)$ represents the probability that an examinee with trait θ_i will "move" from an incorrect response on item j - 1 (state 0) to a correct response on item j (state 1). Likewise, $\beta Q_j(\theta_i)$ represents a transition probability from a correct response on item j - 1 to an incorrect response on item j. The probabilities $1 - \alpha P_j(\theta_i)$ and $1 - \beta Q_j(\theta_i)$ imply state consistency between item responses. It should be noted that items j and j - 1 need not be adjacent items, but may occur anywhere throughout the test.

The parameters α and β are dependency weights where $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$. These dependency weights may take on any value within the specified range and need not be equal. The weights can be assigned values independently of one another, and thus fix the direction of state transition (e.g. increasing or decreasing the probability of going from a correct to an incorrect response). When $\alpha = \beta = 1$, the items are totally independent, and when $\alpha = \beta = 0$, the items are completely dependent.

The probability of answering the jth item correctly, given an incorrect or a correct response to the previous item is defined as $P'_j(\theta_i)$ where:

$$P'_j(\theta_i) = Q'_{j-1}(\theta_i)\, \alpha P_j(\theta_i + P'_{j-1}(\theta_i) \{1 - \beta Q_j(\theta_i)\}$$

and

P'_{j-1} is the probability of a correct response on item j - 1 and

$Q'_{j-1} = 1.0 - P'_{j-1}$.

The degree to which local independence is violated depends upon several factors including the α and β weights, the individual parameters (i.e., difficulty and discrimination) of those items within the dependent sequence, the order of the items (e.g. easy to difficult, difficult to easy), and the length of the dependent sequence. Spray and Ackerman (1986) summarized all of these factors in terms of a statistic ϕ. For a dependent sequence of length m, ϕ is the sum of the absolute differences between the likelihood of each of the possible 2^m response patterns which can occur under a joint density function of the Markov process and the likelihood under an assumption of local independence for the m items.

The absolute value of the differences is evaluated at some value $\theta = \theta_i$ (which is thought to be representative of the entire examinee population) and summed over all possible 2^m response patterns.

Specifically,

$$\phi = \sum_{\ell=1}^{2^m} |P^*(\underset{\sim}{U}_\ell = \underset{\sim}{u}|\theta_0) - P(\underset{\sim}{U}_\ell = \underset{\sim}{u}_\ell|\theta_0)|$$

where

$$P^*(\underset{\sim}{U}_\ell = \underset{\sim}{u}_\ell|\theta_0) = P(U_1 = u_1|\theta_0)P(U_2 = u_2|\theta_0,u_1) \ldots$$
$$P(U_m - u_m|\theta_0, u_{m-1})$$

and

$$P(\underset{\sim}{U}_\ell = \underset{\sim}{u}_\ell|\theta_0) = \prod_{j=1} [P(U_j = u_j|\theta_0)] .$$

It can be shown that for any given ability value, θ_0, $0.0 \le \phi < 2.0$, regardless of whether $P_j(\theta_0)$ is defined as a one, two, or three parameter IRT model. When $\phi = 0.0$, local independence holds throughout the m item sequence. As ϕ increases, the degree of dependency increases.

Although the ϕ statistic is thought to be useful in describing the degree of dependency within a set of items, it is doubtful that it can be recovered in an estimation sense. That is, ϕ is based on all possible response patterns for a "representative" value of θ, while in reality each value of θ may be unique and be represented by only one response pattern which may or may not be the most likely. Thus, ϕ is believed to be more of a tool for describing dependent data rather than an estimable parameter.

In summary, this study had two main objectives. The first objective was to validate the use of the dependency model developed by Ackerman and Spray (1986) as a useful tool for generating dependent data. The second goal was to determine the affect various degrees of dependency had on the IRT calibration process using different sample sizes.

Method

To accomplish these two objectives, response vectors to a calibrated set of item parameters were generated using four levels of dependency with three different sample sizes, producing 12 different data sets. Sets were first examined to insure that four levels of dependency existed in the created response sets. Then each of the 12 sets were calibrated using a two parameter logistic model with the computer programs BILOG and LOGIST. Ability and item estimates were examined to determine how dependency effected the estimation process.

Data Generation

In this study data were generated using four levels of dependence (no dependence, weak, medium, and strong) with three different sample sizes (N = 400, 800, and 1200). The degrees of dependency were selected to be representative of nearly the full range of dependency as described by the ϕ statistic. The specific sample sizes were chosen for three reasons: it was thought that such sample sizes were realistic in terms of academic testing situations; it was felt that the sample size of 400 would be a minimum for use with the two parameter IRT model; and, that differences in calibration accuracy would be most evident in sample sizes having this range.

Data were generated using the ACT Assessment Math Usage Form 26A as a model. It is a 40 item multiple choice test with each item having five alternatives. A brief description of the content of the test is provided in Appendix A.

A dependent sequence length of eight items was chosen because it was thought to be typical of the number of items assigned to a test passage. Items 1-8 were selected to be the dependent block of items. The dependency weights and ϕ values for the four levels of dependency were:

Level of Dependency

	ϕ	α	β
Total independence	0.00	1.00	1.00
Weak	0.66	0.60	0.80
Medium	1.09	0.50	0.40
Strong	1.57	0.30	0.10

Using previous calibrated LOGIST two parameter logistic item parameter estimates, and randomly generated abilities from a N(0, 1) distribution, 1,000 response vectors for the 12 datasets were generated using the dependent model. Each dataset was then calibrated separately using the IRT calibration programs LOGIST and BILOG. The two IRT calibration programs use different estimation procedures. A LOGIST uses a joint maximum likelihood estimation procedures, whereas BILOG uses marginal maximum likelihood estimation. The default method of scoring subjects was selected for all BILOG computer runs. This method of scoring was expectation a posteriori using a normal N(0, 1) Bayesian prior. The default log-normal prior was used in the item discrimination calibration. No prior was used in estimating the difficulty parameters.

Mean inter-item tetrachoric correlations for the dependent and independent items are shown in Table 1. In the total independent datasets the mean correlations ranged from .383 to .433 suggesting a moderate degree of similarity among response patterns for the original 40 items. The effect of the dependency model is clearly demonstrated. As compared to the independent case, the average tetrachoric increases as the level of dependency increases.

After each calibration run, Yen's Q3 statistic (Yen, 1984) was computed and used as a comparison measure to determine how robust the calibration programs were to violations of local independence. Q_3 is defined as the correlation taken over examinees of:

$$d_{i\cdot} = u_{i\cdot} - \hat{P}_i(\hat{\theta}_\cdot) \text{ and}$$

$$d_{j\cdot} = u_{j\cdot} - \hat{P}_j(\hat{\theta}_\cdot)$$

where $u_{i\cdot}$ is the score of an item on item i,

u_{ji} is the score of an item on item j

and $\hat{P}_i(\hat{\theta}_.)$, $\hat{P}_j(\hat{\theta}_.)$ are the probabilities of a correct response based upon the estimated item and ability parameters. A Fisher r-to-z transformation of Q3 is approximately distributed as a normal variable with mean equal to zero and variance equal to 1/(N-3).

Average absolute differences (AAD) between the calibrated items parameters for the independent datasets and the dependent datasets were computed as a measure of item parameter shift.

To examine bias in the estimated abilities, bias $(\theta - \hat{\theta})$ was plotted for each dataset within each calibration program. Correlations between the ability estimates for each level of dependency with equal sample sizes were also calculated.

Results

The mean p value and biserial correlation for items 1-8 and items 9-40 for each level of dependency are shown in Table 2. As the degree of dependency increased the average difficulty (p) increased. This is due in part to the order of difficulty within the dependent block of items. That is, if an easier item precedes and is dependent with a subsequent item, it will have a tendency to make the subsequent item easier, (See Ackerman & Spray, 1986, p. 12-13). Thus, subsequent items tend to become more similar in difficulty to the previous items on which they are dependent. It can also be seen in Table 2 that the items within the dependent block become more homogeneous as the dependency increases. (Note biserial correlations were computed using the item score with the total test score.)

The AAD values and correlations between the estimated item parameters for LOGIST and BILOG are displayed in Tables 3a and b. The AAD discrimination and difficulty values for the dependent items increased dramatically for each estimation program as the level of dependency increased. This was more true of BILOG than LOGIST. AAD difficulty differences did not, however, increase nearly as much as the AAD discrimination differences.

Correlations between b and $\hat{b}$ for the dependent items dropped appreciably from the medium to the strong dependency level. For both LOGIST and BILOG, the $r_{b\hat{b}}$ for the strong dependency case was negative, except for the BILOG calibration when the sample size was 1200. Correlations between the discrimination parameters and their estimates for the dependent items also fell noticeably from the medium to the strong level of dependency. However, the difficulty and discrimination estimates for the independent items appeared to be unaffected by the various levels of induced dependency.

Yen's Q3 analysis of the calibrated parameters is shown in Table 4. The pattern observed by Yen (1984) can also be observed in these data. That is, the more locally dependent the items the higher the Q3 value. By comparing the Q3 computed on the calibrated dependent items data with the Q3 based on the independent item sets one can obtain a relative sense of how robust each calibration program is to the violation of local independence. The Q3 value for the dependent item block increased for each sample size as the dependency level increased. For each level of dependency the Q3 value for the N = 800 case was larger than the two other sample sizes, except for the BILOG estimation of the strong dependency case. This might be a result of sampling variance. However, the overall Q3 values computed using BILOG estimates do not increase as much as though computed using the LOGIST estimated values.

It should be noted that using Bock's χ^2 goodness-of-fit measure (Bock & Mislevy, 1982), each calibration program's item parameter estimates fit the 2PL IRT model for all datasets at the .01 level of significance.

Correlations between estimated abilities for the no dependency case and the three other levels of dependency for each calibration program for each sample size showed little difference. For each sample size as the dependency increased, the correlation between θ and $\hat{\theta}$ dropped from .95 to .90. This suggests that if the violation of local independence is great enough, ability estimation can be affected. No difference was detected for the ability estimates between the two calibration programs.

Bias plots of $\theta - \hat{\theta}$ for each of LOGIST and BILOG calibration runs are shown in Figures 1, 2, and 3. The θ's were rank ordered and divided into quantiles in .2 increments from -3.0 to +3.0. All estimated abilities were rescaled to the θ scale. Mean differences between $\theta - \hat{\theta}$ were obtained for each group and then plotted. As dependency increased both estimation programs appeared to increasingly overestimate abilities at the lower end of the ability scale and underestimate abilities at the upper end of the scale. One possible explanation is the following: Because the dependency was built into the easiest items, low ability students would have a tendency to get more items near their true ability level correct, thus increasing their estimated abilities. Upper ability students would have underestimates of their ability because, although they would be likely to have the same number of correct responses, the items would be less difficult. Less variance in the $\theta - \hat{\theta}$ difference was detected for the sample size of 800, then for N = 400, or N = 1200.

Conclusions

The results of this study have several implications for the calibration of locally dependent items. First, the stronger the violation of local independence (as defined in the model by Ackerman and Spray, 1986) the greater the effect of item parameter calibration regardless of sample size. For cases of strong dependency, item difficulty and discrimination estimates correlated negatively with their parameters. If items of this type are calibrated in a dependent sequence and used separately (e.g., in an adaptive testing pool), the overestimates/underestimates of the parameters could affect ability estimation.

In this study, dependency was injected into an eight item sequence at the beginning of the test where the items were the easiest. Thus, ability estimation was affected more at the low end than the middle and upper levels of the ability scale. If, however, this dependency had been imposed on middle difficulty or very hard items, it is believed that ability estimates at the middle and upper ends of the scale would be more affected than the lower levels of ability.

Very little difference was noted between BILOG and LOGIST's ability estimation. Within each level of sample size, the $\hat{\theta}$ LOGIST and $\hat{\theta}$ BILOG tended to correlate in the neighborhood of .98 to .99. Bias plots of the estimated abilities revealed that both programs are affected as the dependency increases, regardless of sample size. BILOG resolves violations of local independence by overestimating the discrimination parameters more so than LOGIST.

Yen's Q3 statistic validated the Ackerman and Spray dependency model as a useful tool for simulating dependent data. The Q3 results suggest that BILOG is slightly better than LOGIST at calibrating response data in which the assumption of local independence has been violated. Perhaps by imposing appropriate Bayesian priors on the ability and item distribution the BILOG calibration process could be further improved.

These results strongly suggest that the calibration of item parameters should be conducted jointly with a review of the response processes for all items on the calibrated test. If after studying the response process required for each item violations of local independence are suspected, calibration results should be guarded.

This study needs to be replicated to verify the findings. Other directions for future research would be to impose dependency with items of differing difficulty levels to see if there is an effect on ability estimation at other points of the ability scale. Likewise, the results should be replicated with a 3PL model to study the effect of guessing.

REFERENCES

Ackerman, T. A. & Spray, J. A. (1986, April). A general model for item dependency. Paper presented at the AERA Annual Meeting, San Francisco.

Spray, J. A., & Ackerman, T. A. (1986, June). The effect of item response dependency on trait or ability dimensionality. Paper presented at the Psychometric Society Annual Meeting, Toronto.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. BILOG, Maximum likelihood item analysis and test scoring: Logistic model. Scientific Software, Inc. Mooresville, IN.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

TABLE 1

Mean Interitem Tetrachoric Correlations for Items in the Dependent and Independent Item Blocks

ϕ	N	$\bar{r}_{tetra}$ Items 1-8 (Dependent Block)	$\bar{r}_{tetra}$ Items 9-32 (Independent Block)
0.00	400	.433	.488
(no dependency)	800	.383	.477
	1200	.429	.476
0.66	400	.553	.488
(weak)	800	.481	.477
	1200	.527	.476
1.09	400	.644	.488
(medium)	800	.594	.477
	1200	.645	.476
1.57	400	.827	.488
(strong)	800	.855	.477
	1200	.866	.476

TABLE 2

Mean Difficulty, Biserial Correlations, and Reliability Coefficients for the Dependent (D) and Independent (I) Items for each Level of Dependency

	$\bar{p}$		$\bar{r}_{bis}$		
ϕ	D	I	D	I	KR-20
0.0	.61	.36	.66	.73	.94
0.6	.64	.36	.68	.73	.94
1.1	.69	.36	.71	.73	.94
1.5	.79	.36	.79	.72	.94

Note. Sample size was 1200.

TABLE 3a

Correlations and Average Absolute Differences of the Dependent (D) and Independent (I) Items for Values of LOGIST Calibrated $\hat{a}$, $\hat{b}$ and a, b for Different Dependency Conditions

ϕ	N	$r_{a\hat{a}}$		$r_{b\hat{b}}$		$\frac{\sum\lvert a-\hat{a}\rvert}{k}$		$\frac{\sum\lvert b-\hat{b}\rvert}{k}$	
		D	I	D	I	D	I	D	I
0.0	400	.82	.94	.98	.97	.16	.11	.07	.08
	800	.91	.98	.98	.98	.10	.10	.06	.05
	1200	.87	.98	.99	.99	.13	.07	.03	.04
0.6	400	.87	.93	.90	.97	.14	.12	.13	.10
	800	.94	.97	.93	.98	.07	.07	.13	.07
	1200	.85	.93	.93	.99	.11	.07	.12	.05
1.1	400	.60	.94	.82	.97	.37	.12	.23	.08
	800	.50	.96	.76	.98	.23	.13	.25	.06
	1200	.44	.98	.90	.99	.39	.08	.19	.04
1.5	400	-.08	.94	-.17	.97	.95	.17	.33	.08
	800	.02	.98	-.21	.98	1.04	.19	.34	.08
	1200	-.24	.98	-.09	.99	1.25	.13	.31	.05

TABLE 3b

Correlations and Average Absolute Differences of the Dependent (D) and Independent (I) Items between Values of BILOG Calibrated $\hat{a}$, $\hat{b}$ and a, b for Different Dependency Conditions

ϕ	N	$r_{a\hat{a}}$		$r_{b\hat{b}}$		$\frac{\sum\lvert a-\hat{a}\rvert}{k}$		$\frac{\sum\lvert b-\hat{b}\rvert}{k}$	
		D	I	D	I	D	I	D	I
0.0	400	.82	.95	.99	.98	.76	.71	.06	.08
	800	.92	.97	.98	.99	.60	.72	.05	.05
	1200	.87	.99	.99	.99	.74	.70	.03	.04
0.6	400	.86	.94	.90	.98	.74	.58	.13	.09
	800	.92	.98	.93	.99	.48	.58	.13	.06
	1200	.87	.98	.93	.99	.66	.57	.12	.06
1.1	400	.51	.94	.83	.98	1.16	.72	.23	.08
	800	.48	.98	.76	.99	.80	.83	.26	.06
	1200	.43	.98	.91	.99	1.19	.71	.20	.04
1.5	400	-.17	.95	-.21	.98	2.38	.82	.33	.07
	800	-.09	.98	-.23	.98	2.87	.77	.33	.07
	1200	-.15	.98	.17	.99	3.66	.74	.30	.04

TABLE 4

Yen's Q3 Analysis of Dependent and Independent Item Blocks Calibrated with LOGIST and BILOG

		Q3			
		LOGIST		BILOG	
φ	N	DEP	IND	DEP	IND
0.0	400	-.042	-.024	-.032	-.012
	800	-.025	-.024	-.013	-.012
	1200	-.039	-.024	-.018	-.013
0.6	400	.040	-.027	.053	-.015
	800	.053	-.027	.076	-.014
	1200	.040	-.027	.061	-.014
1.1	400	.100	-.029	.106	-.019
	800	.137	-.030	.160	-.021
	1200	.113	-.029	.113	-.020
	400	.267	-.034	.206	-.023
	800	.308	-.036	.192	-.021
1.5	1200	.251	-.033	.124	-.017

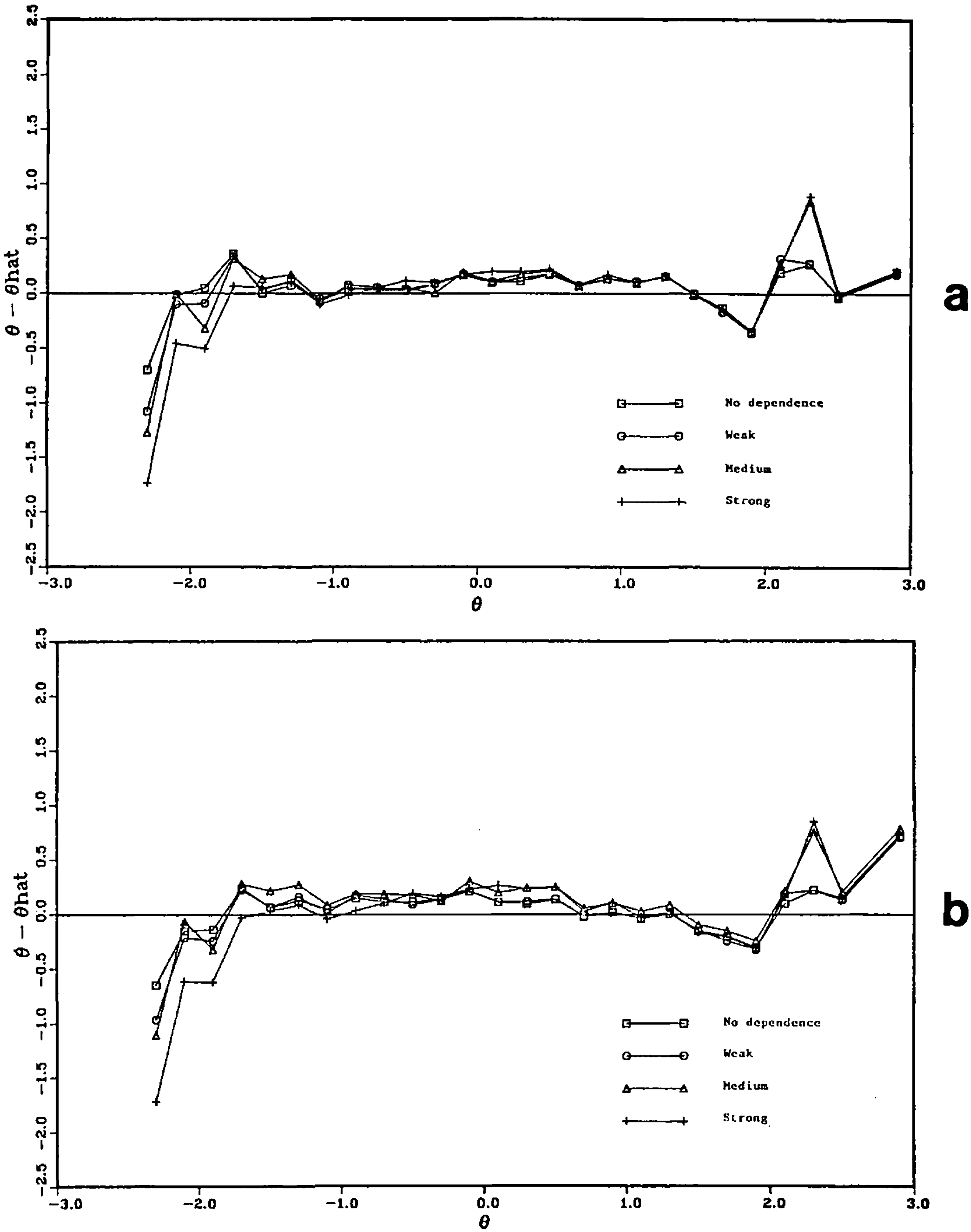

Figure 1. Bias Plots of $\theta - \hat{\theta}$ for LOGIST (a) and BILOG (b) Calibration Runs: N = 400.

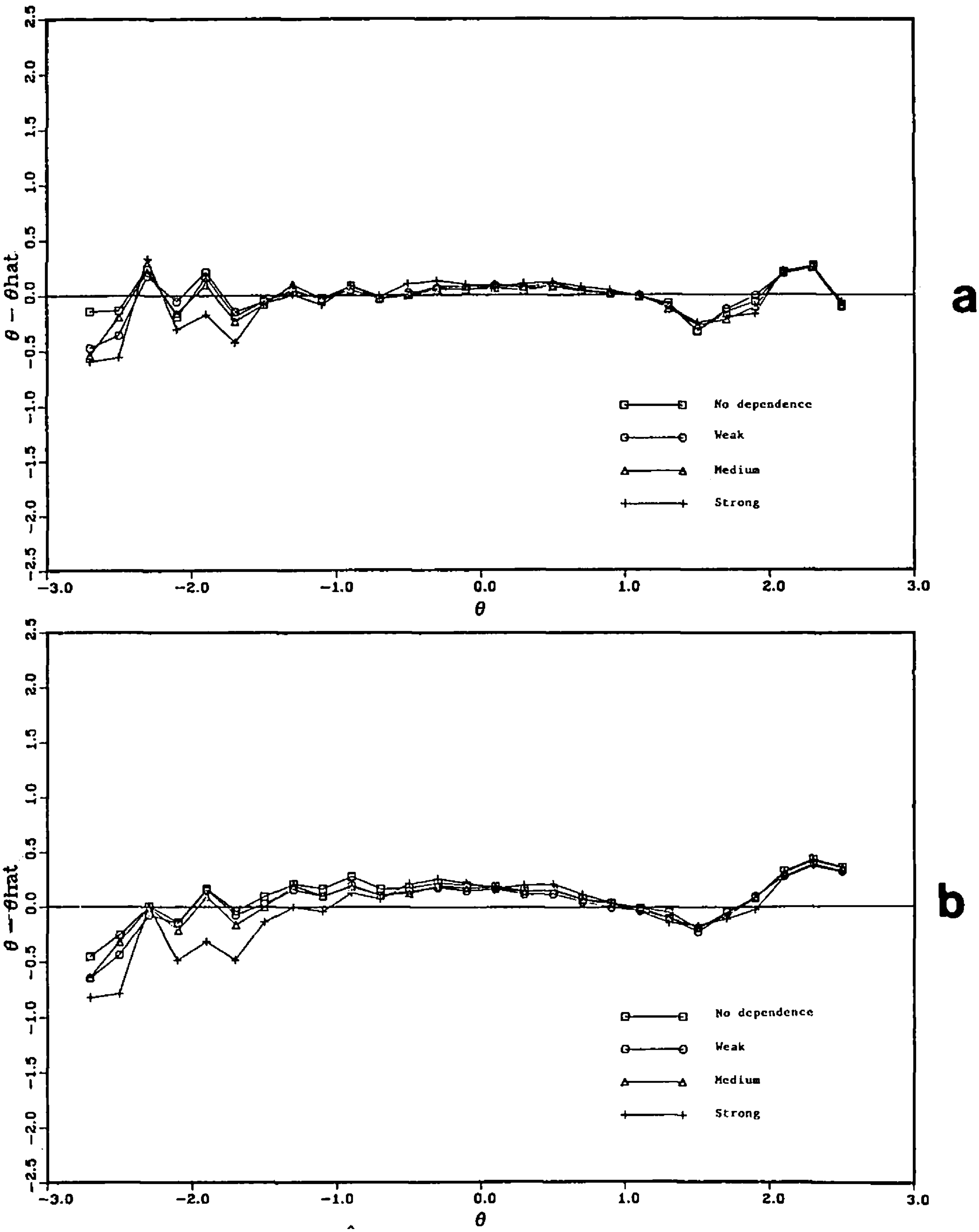

Figure 2. Bias Plots of $\theta - \hat{\theta}$ for LOGIST (a) and BILOG (b) Calibration Runs: N = 800.

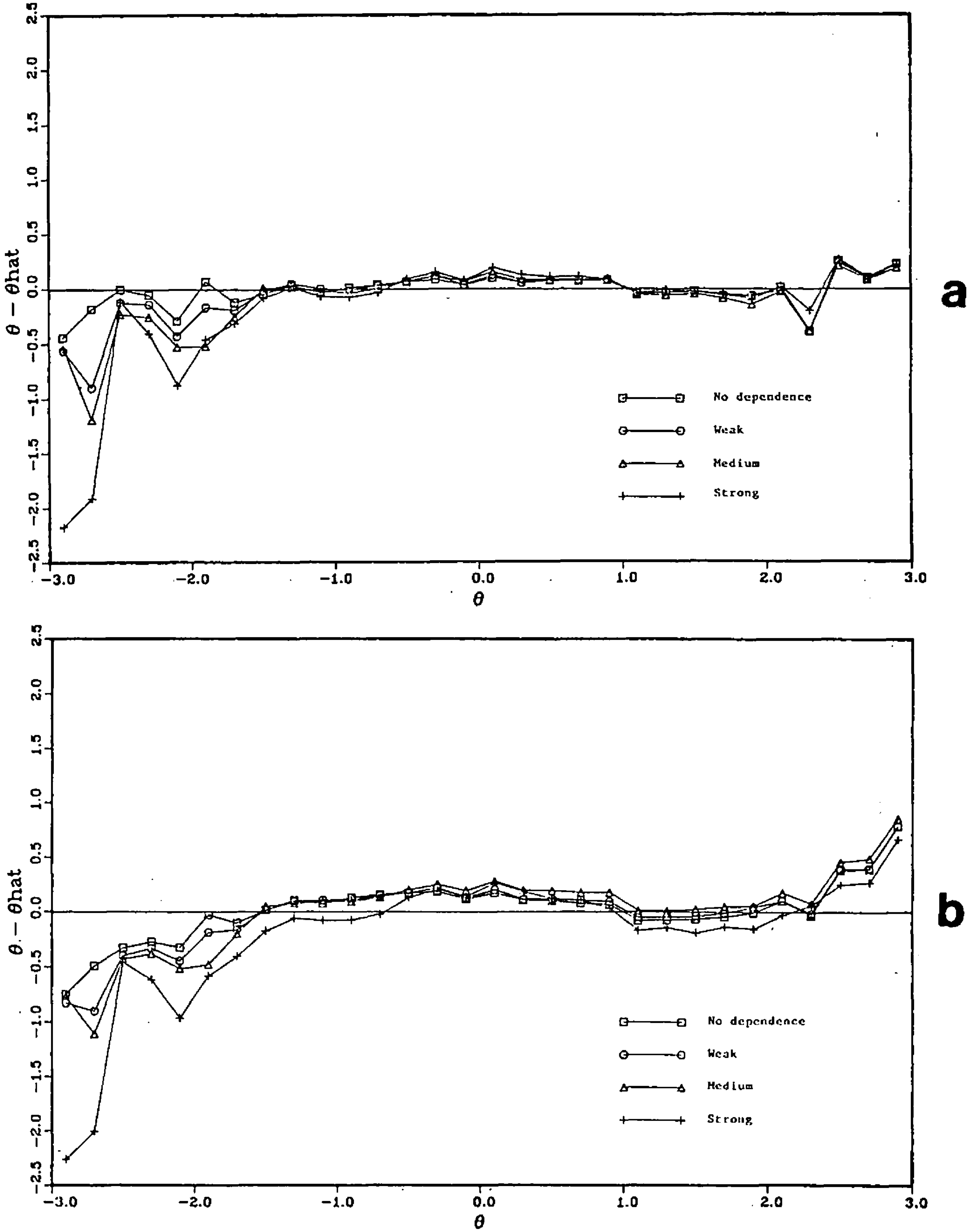

Figure 3. Bias Plots of $\theta - \hat{\theta}$ for LOGIST (a) and BILOG (b) Calibration Runs: N = 1200.

Appendix A

ACT MATHEMATICS USAGE TEST

Description of the test. The Mathematics Usage Test is a 40-item, 50-minute test that measures the students' mathematical reasoning ability. It emphasizes the solution of practical quantitative problems that are encountered in many postsecondary curricula and includes a sampling of mathematical techniques covered in high school courses. The test emphasizes quantitative reasoning, rather than memorization of formulas, knowledge of techniques, or computational skill. Each item in the test poses a question with five alternative answers, the last of which may be "None of the above."

Content of the test. In general, the mathematical skills required for the test involve proficiencies emphasized in high school plane geometry and first- and second-year algebra. Six types of content are included in the test. These categories and the approximate proportion of the test devoted to each are given below.

Mathematics Content Area	Proportion of Test	Number of Items
a. Arithmetic and Algebraic Operations	.10	4
b. Arithmetic and Algebraic Reasoning	.35	14
c. Geometry	.20	8
d. Intermediate Algebra	.20	8
e. Number and Numeration Concepts	.10	4
f. Advanced Topics	.05	2
Total	1.00	40

a. *Arithmetic and Algebraic Operations.* The items in this category explicitly describe operations to be performed by the student. The operations include manipulating and simplifying expressions containing arithmetic or algebraic fractions, performing basic operations in polynomials, solving linear equations in one unknown, and performing operations on signed numbers.

b. *Arithmetic and Algebraic Reasoning.* These word problems present practical situations in which algebraic and/or arithmetic reasoning is required. The problems require the student to interpret the question and either to solve the problem or to find an approach to its solution.

c. *Geometry.* The items in this category cover such topics as measurement of lines and plane surfaces, properties of polygons, the Pythagorean theorem, and relationships involving circles. Both formal and applied problems are included.

d. *Intermediate Algebra.* The items in this category cover such topics as dependence and variation of quantities related by specific formulas, arithmetic and geometric series, simultaneous equations, inequalities, exponents, radicals, graphs of equations, and quadratic equations.

e. *Number and Numeration Concepts.* The items in this category cover such topics as rational and irrational numbers, set properties and operations, scientific notation, prime and composite numbers, numeration systems with bases other than 10, and absolute value.

f. *Advanced Topics.* The items in this category cover such topics as trigonometric functions, permutations and combinations, probability, statistics, and logic. Only simple applications of the skills implied by these topics are tested.