SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

AD-A214 709

1. REPORT NUMBER: AFOSR-TR-89-1654

4. TITLE (and Subtitle)

ESTIMATION OF PARAMETERS IN THE THREE-PARAMETER LATENT TRAIT MODEL

5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report (Feb. 1, 1978-April 30, 1979)

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

Hariharan Swaminathan and Janice A. Gifford

8. CONTRACT OR GRANT NUMBER(s): F49620-78-C-0039

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Laboratory of Psychometric and Evaluative Research
School of Education/University of Massachusetts
Amherst, MA 01003

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Department of the Air Force
Air Force Office of Scientific Research
Bolling Air Force Base, DC 20332

12. REPORT DATE: March 1979

13. NUMBER OF PAGES: 27 pages

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

DTIC ELECTE NOV 29 1989 S B D

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

A paper presented at an AERA-NCME symposium entitled "Explorations of Latent Trait Models as a Means of Solving Practical Measurement Problems," San Francisco, April 1979.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Keywords: Latent trait theory, Cognition, Mathematical models, Psychological tests, Aptitude tests. (SDC)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Two methods for estimation of parameters of the three-parameter logistic model, the Urry method and the maximum likelihood procedure, were studied with respect to several issues using artificial data. Comparisons were made as to the accuracy of estimation and its relationship to the number of items and examinees, the effect of the distributions of ability on the resulting estimates of items and ability parameters, and the statistical properties such as bias and consistency of the resulting estimates.

89 11 27 080

DD FORM 1 JAN 73 1473

Unclassified

Estimation of Parameters in the Three-Parameter Latent Trait Model[1,2,3]

Hariharan Swaminathan and Janice A. Gifford
University of Massachusetts, Amherst

ABSTRACT

Two methods for estimation of parameters of the three-parameter logistic model, the Urry method and the maximum likelihood procedure, were studied with respect to several issues using artificial data. Comparisons were made as to the accuracy of estimation and its relationship to the number of items and examinees, the effect of the distributions of ability on the resulting estimates of items and ability parameters, and the statistical properties such as bias and consistency of the resulting estimates.

---

[1]The project was performed pursuant to a contract from the United States Air Force Office of Scientific Research. However, the opinions expressed here do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred.

[2]Laboratory of Psychometric and Evaluative Research Report No. 90. Amherst, MA: School of Education, University of Massachusetts, 1979.

[3]A paper presented at an AERA-NCME symposium entitled "Explorations of Latent Trait Models as a Means of Solving Practical Measurement Problems," San Francisco, April 1979.

Estimation of Parameters in the Three-Parameter Latent Trait Model

Hariharan Swaminathan
Janice A. Gifford
University of Massachusetts, Amherst

The successful application of latent trait theory to practical measurement problems hinges upon the availability of procedures for the estimation of the parameters. Hence, investigations of the adequacy of the available procedures for estimating parameters in latent trait models are necessary and, indeed, play a crucial role when assessing the usefulness of latent trait theory.

While the problem of estimating parameters in the one-parameter latent trait model appears to be solved, some degree of controversy seems to surround the estimation of parameters in the two- and three-parameter models (Wright, 1977; Andersen, 1973). Lord (1975) has empirically evaluated the maximum likelihood procedure for estimating the parameters in the three-parameter model and has provided answers to some of the questions that arise with respect to estimation of parameters. Jensema (1976) has compared the efficiency of a heuristic procedure suggested by Urry (1974) for estimating the parameters in the three-parameter model with the maximum likelihood procedure. Despite these efforts, little is known regarding the properties of the estimators in the three-parameter model and the effect on the estimates of violating the underlying assumptions, especially with respect to the revised heuristic procedure as suggested by Urry (1976).

The purpose of this study is to investigate the efficiency of the maximum likelihood procedure and the Urry method (Urry, 1976) for estimating parameters in the three-parameter model, to study the properties of the estimators, and to provide some guidelines regarding the conditions under which they should be employed. In particular, the issues investigated are: (1) the "accuracy" of the two estimation procedures, (2) the relationship between the number of items, examinees and the accuracy of estimation, (3) the effect of the distribution of ability on the estimates of item and ability parameters, and, (4) the statistical properties, such as bias and consistency of the estimators.

Design of the Study

In order to investigate the issues mentioned above, artificial data were generated according to the three-parameter logistic model

$$[1] \qquad P_{ij}(\theta) = c_i + (1-c_i)\ \{1 + \exp[-1.7\ a_i(\theta_j - b_i)]\}$$

using the DATGEN program of Hambleton and Rovinelli (1973). Data were generated to simulate various testing situations by varying the test length, the number of examinees, and the ability distribution of the examinees. Test lengths were fixed at 10 items, 15 items, 20 items, and 80 items. Since the accuracy of the maximum likelihood estimation with large numbers of items has been sufficiently documented by Lord (1975), tests with small numbers of items, 10, 15, and 20, were chosen so that the accuracy of the estimation procedure can be ascertained for short tests. This is particularly important if latent trait theory is to be applied to criterion-referenced measurement. Similarly, the sizes of

examinee population were set at 50, 200, and 1000, in order to study the effect of small sample size on the accuracy of estimation.

In the Urry estimation procedure, the relationships that exist for item discrimination and item difficulty between the latent trait theory parameters and the classical item parameters, are exploited (Urry, 1976; Lord & Novick, 1968, pp. 376-378). These relationships are derived under the assumption that the ability is normally distributed and that the item characteristic curve is the normal ogive. In order to study how the departures from the assumption of normally distributed abilities affect the Urry procedure, three ability distributions were considered: the normal, the uniform, and a negatively skewed distribution. The normal and the uniform distributions were generated with mean zero and variance unity (the uniform distribution was generated on the Interval [-1.73, 1.73] to ensure unit variance). A Beta distribution with parameters 5 and 1.5 was generated to simulate a negatively skewed distribution, and then rescaled so that the mean was zero and the variance unity. The distributions were standardized so as to remove the effect of scaling on the estimates of the parameters.

The three factors, test length (4 levels), examinee population size (3 levels), and ability distribution (3 levels) were completely crossed to simulate 36 testing situations. Test data arising from these situations were subjected to the Urry estimation procedure using the computer progams ANCILLES (developed at the U.S. Civil Service Commission) and the maximum likelihood estimation procedure using the computer program LOGIST (Wood, Wingersky & Lord, 1976).

Lord (1975) has emphasized the fact that simulated data should, in some way, resemble real data. Otherwise results obtained through simulation studies will not generalize to real situations. Given this, an attempt was made to generate test data as realistically as possible. In order to accomplish this, item difficulty parameters, b_i, were sampled from a uniform distribution defined on the interval [-2.0, 2.0], and item discrimination parameters, a_i, were sampled from a uniform distribution on the interval [.6, 2.0]. Since data were generated to simulate item responses to multiple choice items with four choices, c_i, the pseudo-chance level parameters, were set at .25. It should be noted, however, that this does not ensure close approximation of the generated data to real data. Combinations of item difficulty and discrimination that may not occur in constructed tests may occur with simulated tests and, hence, affect the estimation procedures, limiting the generalizability of the findings in simulated studies to real situations. On the other hand, since the purpose of this study is to compare two estimation procedures, and to study the statistical properties of estimators, the possible lack of correspondence between simulated and real data may not be a serious problem.

Results

Accuracy of Estimation

Comparisons between the Urry procedure and the maximum likelihood procedure across various test lengths, examinee population sizes, and ability distributions are indicated in Tables 1, 2, and 3. The statistics reported are: (i) the mean, μ, of the population item parameters for each population size, (ii) the mean, $\bar{X}$, of the estimated item parameters, and,

Table 1

Comparison of Estimates of Item and Ability Parameters of the Logist Procedure with the Urry Procedure Based on Normal Distribution of Ability

No. of Items	No. of Examinees	DISCRIMINATION					DIFFICULTY					CHANCE-LEVEL PARAMETER					ABILITY				
			Urry		Logist			Urry		Logist			Urry		Logist			Urry		Logist	
		μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	SD	$\bar{X}$	SD	μ	$\bar{X}$	ρ	$\bar{X}$	ρ
10	50	1.46	3.47	.21	1.53	.43	-.15	-.87	.92	-.60	.95	.25	.34	.38	.12	.04	.02	-.10	.63	.00	.71
	200	1.18	2.82	.08	1.72	.46	.46	.41	.87	.22	.99	.25	.36	.18	.25	.00	.13	.07	.77	-.13	.76
	1000	1.46	2.97	.36	2.00	***	-.15	-.45	.95	-.15	.99	.25	.36	.28	.23	.02	-.00	.11	.71	-.09	.75
15	50	1.17	2.08	-.25	1.67	-.02	.32	.61	.92	.29	.89	.25	.36	.25	.23	.00	.01	.04	.83	-.23	.78
	200	1.17	2.12	.38	1.59	.47	.32	.35	.97	.19	.96	.25	.35	.14	.23	.03	.11	-.00	.77	-.10	.77
	1000	1.40	2.61	.42	1.72	.86	-.09	-.04	.97	-.05	1.00	.25	.33	.17	.25	.00	.02	-.01	.86	-.04	.85
20	50	1.35	2.09	.40	1.60	.37	.16	.22	.95	.04	.96	.25	.29	.14	.18	.04	-.08	-.04	.87	-.00	.87
	200	1.35	2.17	.33	1.41	.46	.16	.22	.97	.08	.97	.25	.30	.12	.24	.01	-.02	.05	.88	-.12	.88
	1000	1.35	1.99	.66	1.59	.76	.16	.37	.98	.16	.99	.25	.36	.11	.25	.02	.00	.05	.89	-.06	.88
80	50	1.28	1.48	.40	1.40	.62	.15	.06	.85	.13	.88	.25	.20	.13	.22	.02	-.08	.12	.96	-.00	.97
	200	1.28	1.42	.64	1.46	.81	.15	.20	.96	.15	.98	.25	.22	.09	.25	.01	-.04	.09	.98	-.00	.97
	1000	1.28	1.36	.84	1.37	.88	.15	.21	.99	.12	1.00	.25	.23	.08	.25	.01	-.00	.08	.98	-.02	.97

Table 2

Comparison of Estimates of Item and Ability Parameters of the Logist Procedure with the Urry Procedure Based on a Skewed Distribution of Ability

No of Items	No. of Exam-inees	DISCRIMINATION					DIFFICULTY					CHANCE-LEVEL PARAMETER					ABILITY				
			Urry		Logist			Urry		Logist			Urry		Logist			Urry		Logist	
		μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	SD	$\bar{X}$	SD	μ	$\bar{X}$	ρ	$\bar{X}$	ρ
10	50	1.18	2.67	.13	1.81	-.38	.46	.79	.91	.68*	.78*	.25	.39	.28	.25	.02	.05	.16	.70	-.30	.78
	200	1.46	2 52	-.58	1.96	-.31	-.15	.31	.98	-.19	.99	.25	.56	.26	.20	.01	-.07	-.01	.71	-.14	.78
	1000	1.46	2.98	-.06	1.95	-.31	-.15	.62	.97	-.2[illegible]	.98	.25	.41	.40	.22	.02	-.01	.03	.57	-.17	.77
15	50	1.40	2.05	.10	1.11	-.01	-.09	.25	.96	-.31	.94	.25	.42	.16	.22	.01	-.17	-.13	.80	-.06	.82
	200	1.40	2.58	.23	1.61	.45	-.09	-.09	.94	-.37	.99	.25	.44	.27	.23	.01	.03	.08	.80	-.12	.91
	1000	1.40	2.37	.18	1.79	.87	-.09	.15	.95	-.10	1.00	.25	.48	.24	.25	.01	.00	.14	.79	-.06	.87
20	50	1.35	2.16	.49	1.25	.27	.16	.28	.92	.16*	.72*	.25	.34	.22	.19	.03	.02	.03	.81	-.14	.89
	200	1.35	2.03	.03	1.54	.10	.16	.21	.96	.05	.98	.25	.41	19	.25	.00	.08	.15	.77	-.05	.87
	1000	1.35	2.10	.49	1.59	.52	.16	.51	.96	.08	.99	.25	.39	.13	.24	.01	.01	.06	.86	-.05	.91
80	50	1.29	1.49	.22	1.19	.28	.18	.10	.85	1.82*	.30*	.25	.21	.16	.21	.01	.07	.22	.93	-.30	.97
	200	1.28	1.26	.69	1.13	.61	.15	.22	.94	1.72*	.27*	.25	.21	.16	.24	.01	.14	.06	.96	-.04	.96
	1000	1.28	1.27	.68	1.24	.82	.15	.38	.97	.16	.99	.25	.20	.11	.25	.01	-.03	.08	.95	-.06	.97

*Indication that the difficulty estimate for an item has taken on an extreme value.

Table 3

Comparison of Estimates of Item and Ability Parameters of the Logist Procedure with the Urry Procedure Based on a Uniform Distribution of Ability

No. of Item	No. of Examinee	DISCRIMINATION					DIFFICULTY					CHANCE LEVEL PARAMETER					ABILITY				
			Urry		Logist			Urry		Logist			Urry		Logist			Urry		Logist	
		μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	SD	$\bar{X}$	SD	μ	$\bar{X}$	ρ	$\bar{X}$	ρ
10	50	1.18	2.50	.33	1.26	.02	.46	.64	.68	.40	.81	.25	.43	.20	.18	.04	.06	.06	.59	.00	.71
	200	1.46	2.86	.60	1.74	.70	-.15	-.28	.90	-.49	.94	.25	.36	.19	.21	.00	.15	.09	.66	-.02	.75
	1000	1.46	2.52	.22	2.00	***	-.15	.06	.98	-.13	.99	.25	.33	.14	.29	.02	-.02	-.04	.74	-.10	.77
15	50	1.40	2.85	.33	1.90	.47	-.09	-.13	.91	-.04	.96	.25	.22	.16	.25	.01	.07	-.04	.90	-.00	.90
	200	1.40	2.70	.13	1.52	.03	-.09	-.04	.92	-.03	.91	.25	.22	.12	.20	.02	-.04	-.04	.89	-.00	.88
	1000	1.40	2.43	.37	1.61	.11	-.09	.16	.95	.20	.87	.25	.25	.09	.23	.01	-.03	-.02	.88	-.03	.87
20	50	1.35	2.35	.09	1.69	.47	.16	.52	.94	.24	.91	.25	.26	.30	.25	.01	-.07	.02	.89	-.14	.88
	200	1.35	2.08	.46	1.68	.34	.16	.40	.92	.07	.98	.25	.35	.24	.25	.00	.09	.05	.91	-.10	.89
	1000	1.35	1.98	.43	1.64	.56	.16	.34	.99	.06	1.00	.25	.46	.34	.24	.02	.04	.03	.90	-.02	.89
80	50	1.29	1.38	.30	1.38	.29	.18	.51	.88	.53	.86	.25	.20	.14	.21	.03	-.30	.09	.95	.00	.96
	200	1.28	1.32	.54	1.38	.73	.15	.29	.93	.20	.95	.25	.30	.15	.23	.01	-.04	.08	.97	.00	.97
	1000	1.28	1.26	.83	1.34	.94	.15	.22	.98	.12	1.00	.25	.36	.18	.25	.00	.02	.08	.97	-.00	.97

(iii) the correlation, ρ, between the true parameters and their estimates. These statistics are reported for both the estimates obtained by employing the Urry procedure and the maximum-likelihood procedure.

A comparison of the mean of the generated item parameters, μ, and the mean of the estimates, $\bar{X}$, for each of the item parameters, discrimination, difficulty, pseudo-chance level and the ability parameters, provides some indication of the accuracy of estimation. However, this comparison is rather weak when carried out alone since the means do not contain all the essential information. Simultaneous comparisons of the means, and examination of the correlations between the parameters and estimates, on the other hand, provide valid information regarding the accuracy of estimation. If the correlation is high, and the means differ, then it can be concluded that the estimation was not sufficiently accurate.

Lord (1975) has implied that if heteroscedasticity exists, it may not be meaningful to compute correlations between true and estimated values. We agree with this, in general. However, since in the strict sense, heteroscedasticity will invalidate the computation of least-squares regression line (the more appropriate criterion to employ is the generalized least-squares criterion), and hence rule out the use of simple, interpretable statistic for the evaluation of the accuracy of estimation, heteroscedasticity (when it occurred) was ignored and correlations and least-squares regression equations were computed.

Estimation of Discrimination Parameter

Examination of the results given in Tables 1, 2, and 3 indicates that the discrimination parameter is poorly estimated for short tests. The highest correlation between true values and estimates for a test with ten items and normally distributed ability is .36, with the mean of the estimates exceeding the mean of the true values. The correlations do improve with increasing sample size and test length, with the mean of the estimated values approaching the mean of the true values from above. The highest correlation between the estimated and true values is .88 for an 80 item test with 1000 examinees. This trend is also evident for uniform and skewed distributions of ability. In general, the discrimination parameter is poorly estimated by the Urry procedure, with the estimation improving more rapidly with increasing test length than with increasing examinee population size.

The least-squares regression lines (for normally distributed ability) for predicting the estimates from true values given in Table 4, were plotted (not shown) and compared with the line y=x, in order to determine the extent of the bias in estimation. The regression lines for all the test length—sample-size combinations fell above the line y=x, indicating that the Urry procedure systematically overestimates the discrimination parameter, with the regression lines approaching the line y=x with increasing test length. Again the "convergence" to the line y=x was more rapid with increasing test length than with increasing sample size.

Trends similar to that observed with the Urry procedure were also observed with the maximum likelihood procedure. Although the estimation of discrimination was poor, the maximum likelihood estimates were consistently better than the "Urry estimates" in that the correlations between

Table 4

Regression Coefficients and Standard Errors for Predicting the Estimates from True Values Based on Normal Distribution of Ability

No. of Items	No. of Examinees	DISCRIMINATION								DIFFICULTY								ABILITY							
		Urry				Logist				Urry				Logist				Urry				Logist			
		b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE
10	50	2.55	1.384	.63	.948	.19	.900	.92	.616	-.71	.168	1.06	.145	.78	.128	1.20	.119	-.11	.165	.48	.113	-.02	.088	.77	.111
	200	2.57	.906	.21	.766	1.17	.359	.45	.273	-.04	.201	.97	.175	-.20	.063	.91	.044	-.01	.058	.59	.036	-.25	.036	.97	.058
	1000	1.52	1.217	.99	.833		***			-.29	.144	1.06	.106	.00	.073	1.03	.043	.11	.025	.58	.019	-.09	.016	.94	.026
15	50	2.89	.800	-.69	.683	1.72	.416	-.04	.352	.30	.143	.98	.106	-.01	.165	.94	.126	.03	.087	.78	.075	-.24	.072	1.00	.116
	200	1.23	.548	.76	.466	.96	.317	.54	.261	.09	.105	.80	.055	-.13	.106	1.01	.072	-.07	.051	.68	.041	-.20	.039	.91	.055
	1000	.85	.991	1.26	.708	.24	.227	1.06	.160	.05	.081	.95	.066	.05	.016	1.08	.000	-.03	.017	.78	.015	-.06	.014	.96	.019
20	50	1.14	.486	.70	.359	.83	.443	.57	.326	.07	.089	.95	.072	-.10	.084	.88	.060	-.10	.078	.82	.068	.07	.072	.87	.070
	200	.70	.988	1.09	.732	.49	.405	.68	.298	.04	.061	1.11	.062	-.09	.059	1.08	.066	.07	.034	.82	.032	.09	.027	1.07	.041
	1000	.36	.406	1.21	.300	.23	.266	1.01	.195	.21	.055	1.02	.045	.00	.035	1.02	.036	.05	.016	.80	.013	-.06	.013	.98	.018
80	50	.46	.267	.80	.208	.26	.162	.89	.125	-.06	.072	.83	.056	-.01	.060	.94	.057	.18	.065	.74	.030	.06	.048	.80	.030
	200	.11	.174	1.02	.135	.23	.104	.96	.079	.05	.035	.97	.033	.04	.031	.74	.024	.12	.018	.89	.011	.04	.020	.93	.016
	1000	.1?	.092	.97	.070	.11	.078	.98	.059	.05	.020	1.03	.015	-.02	.010	.96	.000	.08	.008	.91	.005	-.02	.007	.96	.007

true values and estimates were higher, and the means of the estimates were much closer to the means of the true values. Comparison of the plots of the regression lines given in Table 4 with the line y=x, showed that while there was a general tendency for the parameters to be overestimated, this tendency was not as marked as with the Urry procedure; the "convergence" of the regression lines to the line y=x was more rapid. These trends, higher correlations between true and estimated values than the Urry estimates, tendency for the means of the estimates to be closer to the means of the true values, and rapidity of "convergence" of the regression line to the line y=x, were also observed with the uniform and skewed distribution of ability.

Estimation of Difficulty Parameter

The Urry procedure was extremely successful in providing accurate estimates of the difficulty parameter. The correlations between estimates and true values ranged from .85 to .99. Comparison of the regression lines for normally distributed ability given in Table 4 with the line y=x indicated that except for tests with 10 items, the difficulty parameter was generally overestimated for tests with 15 and 20 items. With larger numbers of items, there was a tendency for difficult items to be overestimated and for easy items to be underestimated. However, the bias was slight in that with increasing items and sample size, the convergence of the regression line to the line y=x was rapid.

The maximum likelihood estimates of the difficulty parameters were, in general, better than the estimates produced by the Urry procedure. The correlations between true and estimated values ranged from .88 to 1.00

(the Urry procedure yielded correlations ranging from .85 to .99). The means of the estimates were, in general, closer to the means of the true values than they were with the Urry procedure. Comparisons of the regression lines, given in Table 4, with the line y=x, revealed that with increasing test length and increasing sample size, the regression line approached the line y=x rather rapidly, demonstrating that there was no bias in the estimation. No clear trends were visible with 10, 15, and 20 items, although the test with 10 items and 50 examinees produced overestimates of the difficulty parameter. These results appeared to hold for both uniform and skewed distributions of ability, although with the skewed distribution there were two instances when the estimates of difficulty went out of bounds. These cases are indicated with an asterisk in Table 2. However, with 80 items and 1000 examinees, the agreement between estimated values and true values was comparable to that obtained with normally distributed ability.

In general, the difficulty parameter was estimated rather well by both maximum likelihood and Urry procedures. The maximum likelihood procedure fared surprisingly well with small numbers of items and examinees in comparison with the Urry procedure, and in general produced better estimates (as determined by the correlations) than the Urry procedure.

Chance-Level Parameter

The true value of the chance-level parameter, c_i, was set at .25 for all the items. Given this lack of variation among the true values, correlations between estimates and true values were not computed. Hence, only the mean of the true values, the mean of the estimates, and the

standard deviation of the estimates are reported in Tables 1, 2, and 3.

The Urry procedure clearly produced very poor estimates of the chance-level parameter. The means of the estimates were consistently higher than the mean of the true values, with relatively large standard deviations. Maximum likelihood estimates, on the other hand, were close to the true values with small standard deviations. The mean maximum likelihood estimates ranged from .12 to .25 for normally distributed ability, from .19 to .25 for skewed distribution of ability, and from .18 to .25 for uniformly distributed ability. In comparison, the Urry procedure yielded estimates that ranged from .20 to .36, .20 to .56, and from .22 to .46, respectively, for the three distributions of ability.

Estimation of Ability

An examination of Tables 1, 2, and 3 indicates a consistent pattern in the estimation of abilities for both maximum likelihood and Urry procedures. The correlations between true values and estimates do not seem to be affected by increasing sample sizes for fixed test lengths. On the other hand, increasing the lengths of the test greatly affect the magnitude of the agreement between true values and estimates. This, not surprising, trend holds for the three distributions of ability.

In general, it appears that although no differences exist between the "Urry estimates" and the maximum likelihood estimates of ability for tests with 15 items or more, the maximum likelihood estimates fare better than the "Urry estimates" for short tests with 10 items. This effect is more pronounced with the skewed ability distribution.

A closer examination of the two estimates carried out by comparing the regression lines, obtained by regressing the estimates on the true

values with the line y=x, indicates that the Urry procedure, in general, underestimates the abilities of examinees with high true abilities and overestimates the abilities of examinees with low true abilities. This may partly be attributed to the fact that the chance-level parameters are overestimated. No such trends were evident with the maximum likelihood estimates. These regression lines rapidly converged to the line y=x with increasing test length.

Effect of Ability Distribution

As pointed out earlier, the Urry procedure exploits the relationships that exist between the classical item parameters and the parameters of the latent trait model. These relationships are derived under the assumption that ability is normally distributed and that the item characteristic curve is the normal ogive. In order to investigate the effect on the estimates of departures from normality, three distributions of ability, the normal, uniform, and a Beta with parameters 5 and 1.5 to simulate a skewed distribution, were generated, and the parameters estimated. A χ^2 test was carried out to determine if the uniform and the Beta distributions deviated sufficiently from the normal. The Beta distribution yielded a χ^2 value of 63.5 when the tails of the normal distribution were excluded and a value of 193.1 when the tails were included. The uniform distribution yielded a χ^2 value of 69.6 when tails were excluded and 307.7 when the tails were included. This indicates that both distributions deviated sufficiently from the normal, with the uniform distribution deviating even more than the Beta distribution.

Comparisons of the results in Tables 1, 2, and 3 reveal that, in general, the Beta distribution affected both estimation procedures, while the uniform distribution produced results similar to those obtained using a normal ability distribution. Although the Beta distribution affected the estimation of discrimination for both procedures, and chance-level and ability for the Urry procedure, the estimation of difficulty did not seem to be affected in either case. The Urry procedure fared poorly with the skewed distribution in comparison to the maximum likelihood procedure in the estimation of the discrimination, chance-level, and ability parameters.

The estimates for the discrimination parameter, resulting from both procedures, were negatively correlated with the true values for short tests. For longer tests, although estimates from both procedures improved, the Urry procedure produced poor estimates in comparison to the maximum likelihood procedure. For an eighty item test with 1000 examinees, a correlation of .68 was obtained using the Urry procedure, as compared to a correlation of .82 obtained from the maximum-likelihood procedure.

The estimates of the chance-level parameters, resulting from the Urry procedure were extremely high for all tests except those of 80 items. The mean values ranged from .20 to .56 with the Beta distribution as compared to a range of .20 to .36 for the normal distribution of ability. The maximum likelihood estimates, on the other hand, were underestimated but comparable to those obtained using a normal distribution of ability.

The maximum likelihood estimates of ability, resulting from using a skewed distribution of ability, were as good as, and in some cases better than, the estimates obtained with a normal distribution. In contrast, the Urry procedure, with a skewed distribution, resulted in poorer estimates. This effect held true even as sample size and test length increased.

In summary, the "Urry estimates" of ability, discrimination, and chance-level parameters seemed to be affected more dramatically than the maximum likelihood estimates,when ability had a skewed distribution. It should be noted that although the uniform distribution had a larger χ^2 value than the Beta distribution, the results obtained with the uniform distribution of ability were similar to those obtained with the normal distribution. It is, then, not departures from normality, but departures from symmetry, and the unavailability of examinees in the lower tail of the ability distribution that affected the estimation procedure.

Statistical Properties of Estimation

Bias. If $\hat{\gamma}$ is an estimator of γ, then g is an unbiased estimator of γ if

$$E(\hat{\gamma}) = \gamma,$$

where $E(\cdot)$ is the expectation operator. This is a desirable property of estimators.

Schmidt (1977) has pointed out that the Urry method based on the procedure developed by Urry (1974) systematically overestimates the discrimination parameter and underestimates the difficulty parameter. Urry (1976) has suggested a correction for this and has incorporated this into the modified Urry procedure employed to estimate parameters in this study. Since it appears that for large numbers of items and examinees the estimates are unbiased (Lord, 1975), in order to study the effect of this correction on the estimates, and to examine if the maximum likelihood estimates are unbiased, a relatively short test (20 items) with 200 examinees was selected, response data generated, item parameters estimated, and replicated 20 times. Since the replications were obtained by generating sets of random examinees, the bias in the estimator of ability was not investigated.

The results of the replications are presented in Table 5 where the true value, μ, of the 20 item parameters are given together with the mean estimate, $\bar{X}$, of the item parameter over 20 replications. The standard error, and the t value obtained as

$$t = (\bar{X} - \mu)/SE$$

are also given to indicate the degree of departure of the mean estimate from the true value.

Table 5

Bias in the Estimation of Item Parameters Based on 20 Replications
(20 Items, 200 Examinees)

	DISCRIMINATION							DIFFICULTY							CHANCE-LEVEL PARAMETER						
			Urry			Logist				Urry			Logist				Urry			Logist	
Item	μ	$\bar{X}$	SE	t	$\bar{X}$	SE	t	μ	$\bar{X}$	SE	t	$\bar{X}$	SE	t	μ	$\bar{X}$	SE	t	$\bar{X}$	SE	t
1	.77	1.17	.067	5.9	.85	.073	1.9	1.63	1.90	.119	2.3	1.65	.097	.2	.25	.37	.020	6.0	.24	.007	-1.4
2	84	1.60	.069	10.9	1.09	.129	1.9	-1.49	-1.28	.079	2.6	-1.52	.107	-.3	.25	.39	.017	8.2	.23	.004	-5.0
3	1.79	2.24	.056	8.04	1.89	.067	1.5	1.82	2.05	.112	2.1	1.75	.005	-1.3	.25	.33	.019	2.1	.23	.006	-3.3
4	1.11	1.91	.054	14.8	1.19	.119	.7	-1.54	-1.57	.076	-.4	-1.91	.188	-2.0	.25	.44	.021	9.0	.23	.004	-5.0
5	1.28	2.35	.202	5.3	1.57	.077	3.8	-.4,	-.38	.206	.4	-.50	.057	-.5	.25	.33	.016	2.2	.23	.004	-5.0
6	1.53	2.42	.151	5.9	1.77	.088	2.6	-1.26	-1.24	.046	.4	-1.30	.063	.6	.25	.35	.036	2.8	.23	.004	-5.0
7	1.31	1.72	.129	3.2	1.62	.105	2.9	1.17	1.33	.079	2.0	1.06	.049	-2.2	.25	.30	.018	2.8	.23	.004	-5.0
8	1.31	1.89	.184	3.2	1.68	.098	3.8	1.47	1.75	.139	2.0	1.37	.067	-1.5	.25	.33	.036	2.2	.23	.005	-4.0
9	1.45	2.48	.117	8.8	1.54	.086	1.0	-1.78	-1.89	.089	-1.2	-1.96	.067	-2.7	.25	.48	.049	4.7	.23	.004	-5.0
10	1.48	2.23	.173	4.3	1.69	.088	2.4	-1.02	-.94	.053	1.5	-1.03	.044	-.2	.25	.34	.028	3.2	.23	.004	-5.0
11	1.58	2.26	.198	3.4	1.78	.081	2.5	.71	.83	.054	2.2	.67	.026	-1.5	.25	.29	.024	1.7	.23	.007	-2.8
12	1.43	1.89	.163	2.8	1.70	.106	2.6	.84	.97	.076	1.7	.74	.034	-2.9	.25	.30	.023	2.2	.23	.006	-3.3
13	1.97	3.06	.192	5.7	1.96	.027	-.4	.19	.07	.062	-1.9	.17	.023	-.9	.25	.18	.030	-2.3	.22	.007	-4.3
14	1.52	2.54	.133	7.7	1.62	.104	.9	-1.64	-1.72	.629	-.1	-1.89	.059	-4.2	.25	.41	.042	3.8	.23	.004	-5.0
15	.73	1.24	.105	4.9	.93	.112	1.8	.01	.22	.059	3.6	.01	.043	0.0	.25	.34	.019	4.7	.23	.006	-3.3
16	1.49	2.07	.221	2.7	1.75	.089	2.9	1.07	1.21	.073	1.9	1.03	.045	-.9	.25	.28	.026	1.2	.23	.005	-4.0
17	1.15	2.05	.129	6.9	1.47	.108	2.9	-.25	-.04	.045	4.7	-.23	.029	.7	.25	.35	.024	4.2	.22	.006	-5.0
18	1.89	2.29	.237	1.7	1.77	.089	-1.4	1.53	1.86	.135	2.4	1.57	.068	.6	.25	.31	.035	1.7	.23	.007	-1.4
19	1.23	1.58	.096	3.6	1.41	.127	1.4	1.28	1.77	.125	3.9	1.31	.086	.3	.25	.36	.016	6.9	.23	.004	-5.0
20	1.20	1.71	.076	6.7	1.56	.094	3.8	.94	1.05	.589	.2	.81	.068	-1.9	.25	.33	.010	8.0	.23	.007	-2.8

The Urry procedure clearly overestimates the discrimination parameter as does the maximum likelihood procedure. However, the bias in the maximum likelihood estimates does not appear to be as severe as the bias in the Urry estimates. This finding is borne out in Figure 1 where the regression line for predicting $\bar{X}$ from μ is plotted for both Urry and maximum likelihood procedures and compared with the line y=x. The maximum likelihood regression line is closer to the line y=x and shows that small values of discrimination are overestimated while very large values tend to be estimated accurately, partly due to the fact that an upper limit was imposed on the estimates. On the other hand, the Urry procedure tends to overestimate large values even more than small values of discrimination.

With item difficulty, the maximum likelihood procedure tends to underestimate easy items, while producing relatively accurate estimates of very difficult items (Figure 2). The Urry procedure, on the other hand, tends to overestimate items with large difficulty levels and underestimate items with negative difficulty levels. In general, the Urry procedure seems to produce biased estimates of item difficulty throughout the entire range.

Consistency. If g_n is an estimator of γ, g_n is a consistent estimator of γ if for any positive ε and η there is some N such that

$$\text{Prob}\{|g_n - \gamma| < \varepsilon\} > 1-\eta,\ n > N.$$

Consistency is a desirable property in that it ensures that an estimator tends to a definite quantity which is the true value to be estimated.

The problem of consistency has raised several questions concerning the estimation of parameters in the latent trait models. Andersen (1972) has argued that a consistent estimator of the discrimination parameter does not

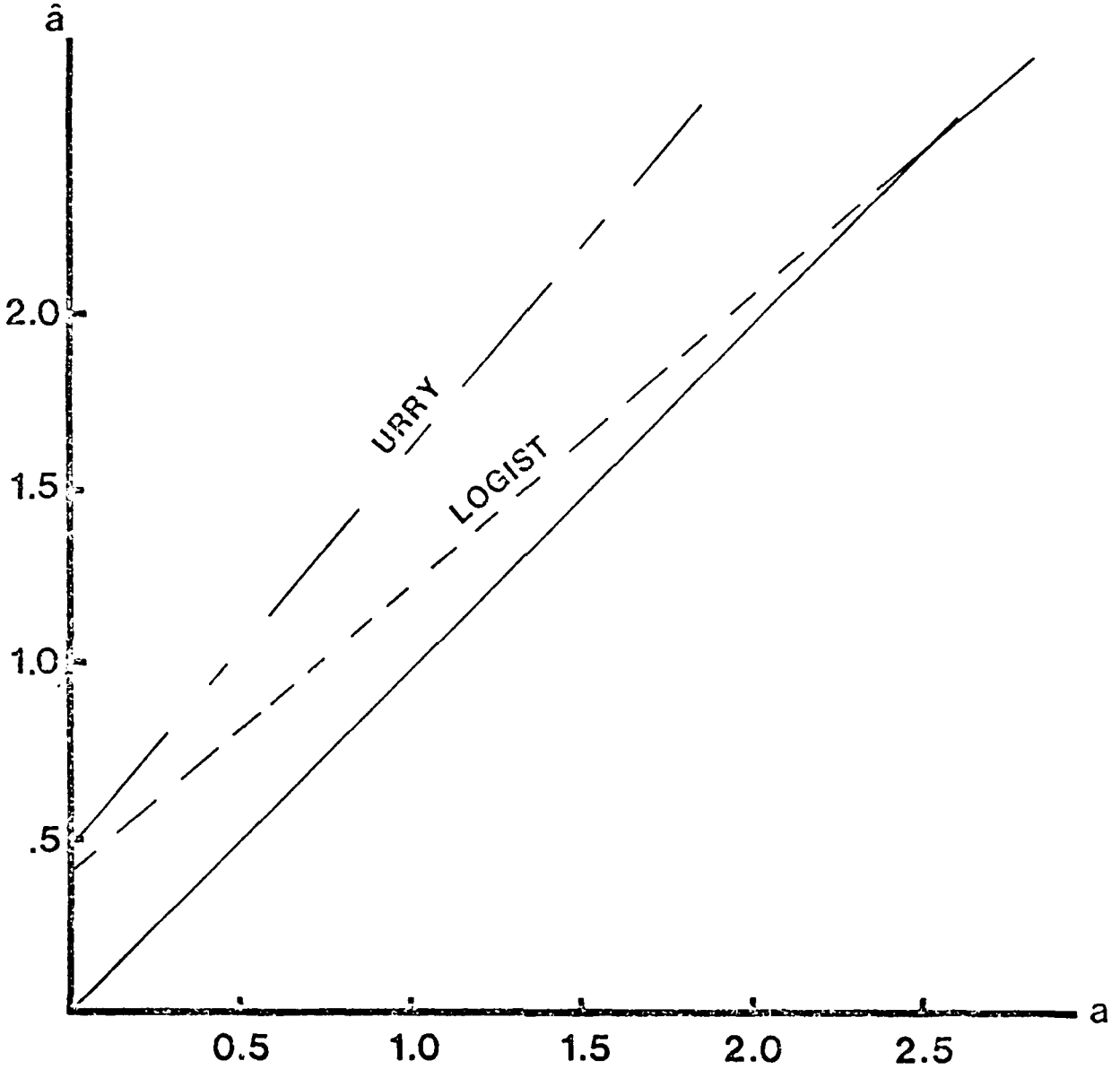

Figure 1. Bias in the estimation of the discrimination parameter.

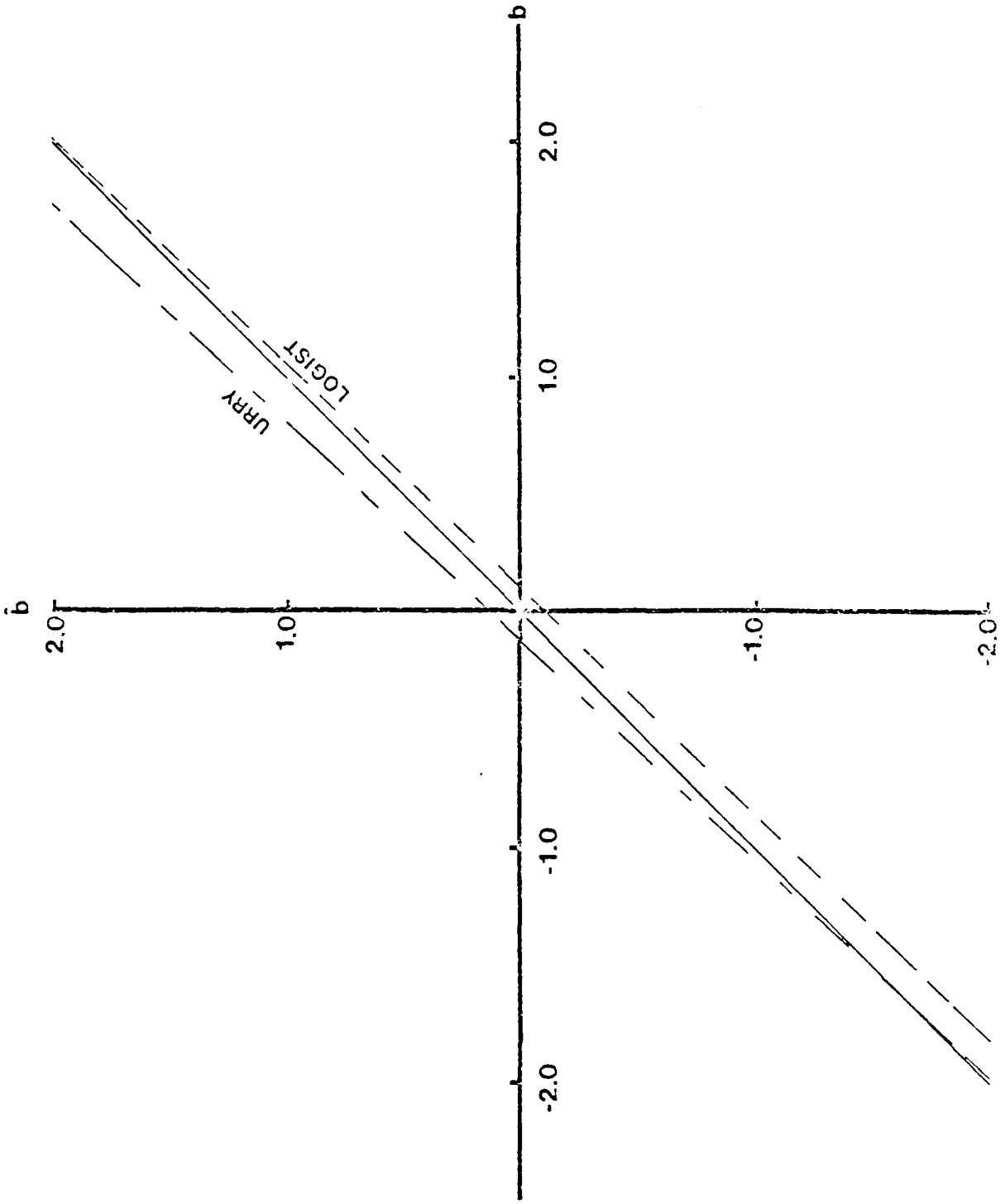

Figure 2. Bias in the estimation of the difficulty parameter.

exist and hence has questioned the meaningfulness of the two- and three-parameter models.

In order to investigate whether or not the maximum likelihood estimators and the "Urry estimators" are consistent, the regression equation for predicting the estimates from the true values of the various parameters were examined. Since the definition for a consistent estimator given earlier implies that an estimator is consistent if (i) it is asympotically unbiased, and (ii) its variance tends to zero with increasing sample size, in order for the estimators of the latent trait parameters to be consistent, (i) the slope of the regression equation must approach one and the intercept approach zero, (ii) the variance, and hence, the standard errors of the estimate of the slope and intercept must approach zero. If these conditions are met then the estimator is consistent.

The regression coefficients and the standard errors are reported in Table 4. The results reported indicate that when both the number of items and the number of examinees increase, the slope and intercept coefficients approach one and zero respectively, with the standard errors approaching zero. This tendency is evident for both Urry and maximum likelihood estimators for the discrimination parameter, difficulty parameter, chance-level parameter and the ability parameter. In all these cases, the maximum likelihood estimator converges in probability to the true value more rapidly than the Urry estimator. It should be pointed out, however, that the results reported here do not conclusively support this. It is clearly necessary to examine the standard errors and the regression coefficients with a greater number of items and examinees.

DISCUSSION

The purpose of this study was to compare two methods for estimation of parameters in the three-parameter logistic model, the Urry method of estimation and the maximum likelihood procedure. The computer programs that were used to carry out this study were the ANCILLES program and the LOGIST program (Wood, Wingersky, & Lord, 1976). The efficiency of the procedures were compared with respect to the accuracy of estimation, the effect of violating underlying assumptions (for the Urry procedure), and the statistical properties of the estimators. The factors that were controlled were: test length (4 levels), examinee population size (3 levels) and ability distribution (3 levels).

The results indicate that, in general, the maximum likelihood procedure is superior to the Urry procedure with respect to the estimation of all item and ability parameters. The differences were pronounced in the estimation of the discrimination and chance-level parameters, while with respect to the estimation of ability and difficulty parameters, the differences were less remarkable. Differing ability distributions had little effect on the estimation of difficulty and ability parameters. However, with a skewed distribution of ability, the Urry procedure produced poorer estimates of discrimination and chance-level parameters than with normal or uniform ability distributions. The maximum likelihood procedure, although faring better than the Urry procedure (with the exception of the 10 item test), produced slightly poorer results with the skewed distribution than the normal or uniform distribution.

The number of examinees had a slight effect in improving the accuracy of estimation of the difficulty, and the chance-level and ability

parameters. However, increasing the number of items and the number of examinees considerably improved the accuracy of the discrimination estimates with both procedures. Surprisingly enough, a twenty-item test with 1000 examinees produced excellent estimates of the difficulty and chance-level parameters, and reasonably good estimates of the discrimination and ability parameters. Tests with 80 items and 1000 people fared considerably better, providing good estimates of all parameters. Testswith 15 items or less while yielding good estimates of difficulty and chance-level parameters, and reasonable estimates of ability parameters, yielded poor estimates of the discrimination parameter. This severely limits the application of the three parameter latent trait model to criterion-referenced measurement situations since criterion-referenced tests typically have fewer than 10 items. However, it should be pointed out that this limitation exists only if the item parameters and ability parameters are estimated simultaneously. If item banks with known item characteristics are employed to estimate ability, or if the Rasch model is employed, this limitation may not exist.

Although the maximum likelihood estimates were superior to the Urry estimates, especially in the case of short tests, the difference between them was negligible when the number of items and the number of examinees increased. This is of particular importance, since the Urry procedure requires considerably less computer time than the maximum likelihood procedure. The time taken for the maximum likelihood procedure, especially with large numbers of items and examinees may become forbidding enough to warrant the use of Urry procedure in this situation. It should be noted, in fairness to the maximum likelihood procedure, the Urry

procedure, in general, deletes more items and examinees during the estimation than the maximum likelihood procedure. This may explain the rapidity of convergence and indicate a weakness in the Urry procedure.

The bias and consistency results indicate that for small numbers of items, the estimates of the item and ability parameters are biased, with the Urry estimates being more biased than the maximum likelihood estimates. As the number of examinees and the number of items increase, it appears that the estimators are unbiased, and in fact, are consistent. This in a sense supports a conjecture of Lord (1968) and shows that the three-parameter model may be statistically viable.

References

Andersen, E. B. Conditional inference in multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 1973, 26, 31-44.

Hambleton, R. K., & Rovinelli, R. A FORTRAN IV program for generating examinee response data from logistic test models. Behavioral Science, 1973, 18, 74.

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

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. Research Bulletin 75-33, Princeton, N.J.: Educational Testing Service, 1975.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

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

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

Urry, V. W. Ancillary estimators for the item parameters of mental tests. Washington, DC: Personnel Research and Development Center, U.S. Civil Service Commission, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976 (revised 1978).

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.