DTIC FILE COPY

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

AD-A214 712

1. REPORT NUMBER: AFOSR-TR-89-1653	2. GOVT ACCESSION NO.
4. TITLE (and Subtitle): EFFECTS OF TEST LENGTH AND SAMPLE SIZE ON THE ESTIMATES OF PRECISION OF LATENT ABILITY SCORES	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report (Feb. 1, 1978-April 30, 1979)
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Linda L. Cook and Ronald K. Hambleton	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: 33 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)

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)

89 11 27 059

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

One of the most important advantages that accrue from the application of latent trait models is the possibility of specifying a target information curve and then selecting items from an item pool to produce a test with the features characterized by this curve. By proceeding in this manner, it is possible to develop a test that provides a pre-specified level of precision (Standard Error of Ability Estimate) at selected ability levels. One problem with this paradigm is that little is known about the precision of the standard error of ability estimates (SEE) under varying circumstances.

DD FORM 1 JAN 73 1473

Unclassified

AFOSR-TR- 89- [redacted]
1653

20. Abstract (continued)

The purpose of the research reported in this paper was to address three practical questions of importance and interest to test developers:

1) What are the effects of examinee sample size and test length on the precision of SEE Curves?

2) What effects do the statistical characteristics of an item pool have on the precision of SEE Curves? and

3) What is the relationship between test length and SEE Curves in typical item pools? Keywords: Latent trait theory,

The results of this study indicated that: (1) both test length and sample size are extremely important factors in the precision of SEE Curves, (2) the precision of SEE Curves at the extremes of an ability continuum would be acceptable in most cases if the curves are based on 200 or more examinees with tests with at least 20 items and, (4) the most sizable improvements in the precision of SEE Curves occur when examinee sample size is increased from 50 to 200 and when test length is increased from 10 to 20 items.

Psychological tests, Aptitude tests, Mathematical models. (SDW)

3/30/79

Effects of Test Length and Sample Size on the Estimates of Precision of Latent Ability Scores[1,2,3,4]

Linda L. Cook
Educational Testing Service

and

Ronald K. Hambleton
University of Massachusetts, Amherst

[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. 87. 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.

[4]The authors are indebted to Janice Gifford for her extensive help in the collection and analysis of data reported in the paper.

There have been a number of highly successful applications of latent trait models in the last couple of years. Reviews of many of these applications are provided by Hambleton, Swaminathan, Cook, Eignor, and Gifford (1979), Rentz and Rentz (1978), and Weiss (1978). The one-, two-, and three-parameter logistic latent trait models have been used by measurement specialists to solve problems in the areas of tailored testing (Weiss, 1978), test score equating (Lord, 1977, in press; Marco, 1977; Rentz & Bashaw, 1977) test development (Wright & Stone, 1978), and item bias (Lord, in press). In fact, the applications cited, and others, have been so successful that the discussionsabout the use of latent trait models have shifted from a consideration of the potential of latent trait models relative to classical models, to a consideration of (1) latent trait models which should be used with particular measurement problems and (2) technical problems (e.g., parameter estimation and goodness of fit measures) arising in connection with the application of particular latent trait models.

This paper was prepared to report some of our recent work in using the three-parameter logistic model in test development. One of the features of using any latent trait model is the possibility of specifying a "target information curve" and then selecting test items from an item pool to produce a test with the features characterized by the "target information curve." A target information curve describes the desired level of "information" at each point on the ability scale underlying examinee test performance. Information, in turn, is directly related to the degree of precision of ability estimates at different points on the ability continuum. In fact, as long as a test is not too short,

the standard error of estimation at a particular ability level is equal to one divided by the square root of information provided by the test at the ability level in question (SEE (θ) = $1/\sqrt{\text{information }(\theta)}$). In practice, since the contribution of each test item to the test information curve (referred to as a "score information curve" when item parameter estimates are used instead of the item parameter values) is known (once the item parameter values or the item parameter estimates are specified), it is possible to select test items from a pool of "calibrated" test items (i.e., a pool of test items with associated parameter estimates) to produce a "score information curve" which approximates a desired "target information curve." With the three-parameter logistic model, items are described by three parameters, referred to as "item difficulty," "item discrimination," and "item pseudo-chance level" (Hambleton et al., 1979).

One of the problems with the paradigm offered above for test development is the imprecision associated with the item parameter estimates. Score information curves (and therefore the associated standard errors of ability estimates) will depend on the precision of item parameter estimates. In turn, precision of item parameter estimates is influenced by the examinee sample size used to estimate the item parameters, and in the case of the item discrimination parameter, estimates are influenced by the length of the test. This study was designed to address three practical questions which are of some importance and interest to test developers:

1. What are the effects of examinee sample size and test length on the precision of standard error of ability estimation curves?

2. What effects do the statistical characteristics of an item pool have on the precision of standard error of ability estimation curves?

3. What is the relationship between test length and standard error of ability estimation curves in typical item pools?

A computer simulation study was chosen as the mode of investigation for the three questions because of the large number of variables which were to be studied, and the need to "know" in some instances, the values of the item parameters.

The remainder of the paper is divided into four sections: (1) Background on Item and Score Information Curves, (2) Method of Investigation, (3) Results, and (4) Conclusions.

Background on Item and Score Information Curves

Once a latent-trait model is specified, the precision with which it estimates examinee ability can be determined. Birnbaum (1968) defined the notion of information as a quantity inversely proportional to the squared length of the confidence interval around an estimate of an examinee's ability. The standard error of ability estimation is equal to $1/\sqrt{\text{information}}$. When information at an ability level is high, narrow confidence bands around the estimates result. If information is low, wider confidence bands are obtained. Because the test information curve varies with ability level, it has been suggested that test information curves ought to replace the use of classical reliability estimates and standard errors of measurement in test score interpretations.

In mathematical terms, Lord (in press) gives the test information curve by

$$I(\theta) = \sum_{g=1}^{n} \frac{P'^2_g}{P_g Q_g} \qquad [1]$$

and the standard error of estimation curve by

$$SEE(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad . \qquad [2]$$

In the expressions above, $I(\theta)$ is the amount of information at ability level θ, $SEE(\theta)$ is the degree of precision of an ability estimate at ability level θ, P_g is the probability of a correct answer to item g by an examinee with ability level θ; Q_g is equal to $1-P_g$; and P'_g is the slope of the item characteristic curve at ability level θ. When item parameter estimates are used in Equation [1], Lord (in press) substitutes the term "score information curve" for "test information curve."

The quantity P'^2_g/P_gQ_g is the contribution of item g to the information curve of the test and is referred to as the item information curve. Item information curves have an important role in determining the accuracy with which ability is estimated at different levels of θ. Each item information curve depends on the slope of the particular item characteristic curve and the conditional variance of test scores at each ability level θ. The higher the slope of the item characteristic curve and the smaller the conditional variance, the higher will be the item information curve at that particular ability level. The height of the item information curve at a particular ability level is a direct measure of the usefulness of the item for precisely measuring ability at that level.

Method of Investigation

Description of the Variables

(a) Test Length

Tests of three lengths were considered: 10, 20, and 80 items. A test with 10 items is about as short a test as is used in practice and therefore the 10-test item length was studied. An 80-item test was considered because the length represents about as long a test as is used in practice.

(b) Ability Distribution

In this particular study, ability scores were simulated to be normally distributed (mean = 0, sd = 1). This assumption was made to conform with a very important assumption made in the item parameter estimation method selected for the study (Urry, 1974). Actually, the parameter estimation method used is a slight modification of the one Urry reported in his 1974 paper. He refers to this new method as "ancillary estimation method." Urry's method was chosen for the study because (1) the method has been extensively used and found to give acceptable results and (2) Urry's computer program is inexpensive.

(c) Sample Size

Three examinee sample sizes were chosen: 50, 200, and 1000. The smallest sample size (N=50) is considerably smaller than anyone should use in practice. It was chosen to identify the "worst possible" results that could be expected. The other two sample sizes define minimum and maximum sample sizes typically used in test development work with latent trait models.

(d) Item Pools

Ranges of parameter values for items in the two pools are shown below:

Item Parameter	Range of Values Pool One	Pool Two
Difficulty (b)	-2.00 to 2.00	-1.00 to 1.00
Discrimination (a)	.60 to 2.00	.60 to 1.50
Pseudo-Chance (c)	.25 to .25	.25 to .25

The differences between the two item pools can be described as follows: Items in pool one had a wider range of difficulty and discrimination values.

Simulation of Data

The eight steps in the simulation study were as follows:

1. Item pool one was selected for study.

2. A test length (10, 20, or 80 items) and a sample size (50, 200, or 1000 examinees) were selected. A sample of examinee ability scores were drawn from a normal distribution (mean=0, sd=1).

3. Using a computer program, DATAGEN (Hambleton & Rovinelli, 1973), (1) item parameters, given the constraints of the item pool under investigation, and (2) examinee item scores were produced. The computer program assumed the correctness of the three-parameter logistic model, used the ability scores from step 2 and item parameters generated at this step, to produce probabilities of correct answers for examinees to the test items. These probabilities, in turn, were converted to examinee item scores (0 or 1) via the use of a random number generator.

4. The examinee item scores from step 3 were used in Urry's computer program to estimate item and ability parameters. However, only the item parameter estimates were used further in this particular study.

5. The item parameter estimates were used in Equation 2 to obtain SEE(θ). The value of SEE(θ) at seven ability levels (θ = -3.00, -2.00, -1.00, 0.00, 1.00, 2.00, 3.00) was calculated.

6. Steps 3 to 5 were repeated three times to obtain three estimates of SEE(θ). All item and ability parameter values for the three runs were identical. The particular examinee item scores varied from one run to the next because of the probabilistic nature of the score outcomes.

7. Steps 3 to 6 were repeated for each combination of test length and sample size (3x3=9).

8. Steps 2 to 7 were repeated with the second item pool. In all, 54 sets of test data were considered in the study.

Results

Effects of Sample Size and Test Length on the Precision of Standard Error of Ability Estimation Curves

In the remainder of this paper "Standard Error of Ability Estimation Curves" will be referred to as "SEE Curves" for convenience.

Tables 1 to 6 contain the SEE Curves with Item Pool One obtained for three replications of three examinee sample sizes (N=50. 200, 1000) and three test lengths (n=10, 20, 80) and reported for seven ability levels. Table 1 to 3 and 4 to 6 contain the same information. What differs is the way the data are organized in the two sets of Tables. Data have been arranged in Tables 1-3 to facilitate an examination of the effect of sample size on SEE Curves. The data presented in Tables 4-6 have been arranged to facilitate an examination of the effect of test length on SEE Curves. Test lengths and sample sizes given under the column headed "actual" are the number of items and examinees remaining after a satisfactory set of item and ability parameter estimates are obtained from Urry's computer program.

Table 1

Summary of Standard Error Estimates[1] for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool (Test Length = 10 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	10	34	.66	.33	.67	.22	.75	1.60	2.19
	2	10	34	2.40	1.88	.56	1.04	.20	1.34	1.37
	3	9	34	.73	.57	1.03	.22	.58	.43	2.19
200	1	10	172	.64	.21	.52	2.15	1.60	1.50	1.48
	2	10	137	.22	.51	.36	1.30	.37	.96	2.45
	3	10	174	2.63	2.14	.27	2.75	.92	.76	1.91
1000	1	10	841	.98	.26	.58	1.43	3.33	.57	1.18
	2	10	833	1.03	1.03	.67	1.05	.45	1.01	1.06
	3	10	892	2.44	.49	.67	.30	.29	.89	1.33

[1]All estimates have been adjusted to correspond to 10-item tests.

Table 2

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool (Test Length = 20 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	20	50	2.84	.70	.35	.30	.31	.44	1.23
	2	20	50	1.93	1.53	.39	.32	.24	.45	1.19
	3	20	46	2.07	.83	.58	.31	.36	.68	1.48
200	1	20	193	--	.57	.26	.39	.33	.50	.77
	2	20	196	--	1.51	.37	.34	.25	.53	.86
	3	20	196	--	1.03	.22	.49	.34	.40	1.15
1000	1	20	955	--	1.05	.48	.33	.33	.45	.82
	2	20	969	--	1.18	.37	.33	.37	.40	.99
	3	20	968	--	1.56	.40	.42	.32	.43	1.07

Table 3

Summary of Standard Error Estimates[1] for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool (Test Length = 80 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	74	50	1.10	.35	.14	.14	.24	.24	.45
	2	79	50	1.06	.48	.25	.17	.13	.32	.49
	3	77	50	.93	.20	.19	.15	.17	.29	.48
200	1	80	200	.89	.26	.22	.24	.19	.25	.44
	2	80	200	.62	.29	.25	.19	.21	.25	.46
	3	80	200	1.06	.35	.21	.19	.20	.25	.48
1000	1	80	999	1.00	.35	.23	.21	.21	.24	.40
	2	80	1000	.98	.32	.23	.22	.21	.23	.43
	3	80	1000	1.08	.34	.20	.21	.20	.24	.46

[1]All estimates have been adjusted to correspond to 80-item tests.

Table 4

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Heterogeneous Item Pool (Sample Size = 50 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	34	.66	.33	.67	.22	.75	1.60	2.19
	2	10	34	2.40	1.88	.56	1.04	.20	1.34	1.37
	3	9	34	.73	.57	1.03	.22	.58	.43	2.19
20	1	20	50	2.84	.70	.35	.30	.31	.44	1.23
	2	20	50	1.93	1.53	.39	.32	.24	.45	1.19
	3	20	46	2.07	.83	.58	.31	.36	.68	1.48
80	1	74	50	1.10	.35	.14	.14	.24	.24	.45
	2	79	50	1.06	.48	.25	.17	.13	.32	.49
	3	77	50	.93	.20	.19	.15	.17	.29	.48

Table 5

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Heterogeneous Item Pool (Sample Size = 200 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	172	.64	.21	.52	2.15	1.60	1.50	1.48
	2	10	137	.22	.51	.36	1.30	.37	.96	2.45
	3	10	174	2.63	2.14	.27	2.75	.92	.76	1.91
20	1	20	193	--	.57	.26	.39	.33	.50	.77
	2	20	196	--	1.51	.37	.34	.25	.53	.86
	3	20	196	--	1.03	.22	.49	.34	.40	1.15
80	1	80	200	.89	.26	.22	.24	.19	.25	.44
	2	80	200	.62	.29	.25	.19	.21	.25	.46
	3	80	200	1.06	.35	.21	.19	.20	.25	.48

Table 6

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Heterogeneous Item Pool (Sample Size = 1000 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	841	.98	.26	.58	1.43	3.33	.57	1.18
	2	10	833	1.03	1.03	.67	1.05	.45	1.01	1.06
	3	10	892	2.44	.49	.67	.30	.29	.89	1.33
20	1	20	955	--	1.05	.48	.33	.33	.45	.82
	2	20	969	--	1.18	.37	.33	.37	.40	.99
	3	20	968	--	1.56	.40	.42	.32	.43	1.07
80	1	80	999	1.00	.35	.23	.21	.21	.24	.40
	2	80	1000	.98	.32	.23	.22	.21	.23	.43
	3	80	1000	1.08	.34	.20	.21	.20	.24	.46

For ease of interpretation, the same data reported in Tables 1 to 6 is presented in graphical form in Figure 1.

Tables 7 to 12 contain similar data to Tables 1 to 6. Tables 7 to 12 contain SEE Curves with Item Pool Two. (There is no figure, however, corresponding to Figure 1 for Item Pool Two.) Tables 13 and 14 were constructed to organize the data reported in Tables 1 to 12 to facilitate the interpretation of results.

(a) Item Pool One—Effect of Sample Size

The results of the simulations for a fixed test length of 10 items, which are reported in Table 1, clearly show the lack of stability of the SEE Curves for all sample sizes. There was little improvement, if any, due to increasing sample size. This result, however, may be due to the limited amount of data considered since improvements were obtained in Item Pool Two and at other test lengths.

From examination of Table 2, which contains the results of the 20 item simulations, it is apparent that the SEE Curves were beginning to stabilize. Except at extreme values of the ability continuum the results were nearly as good as those obtained with the larger sample size (N=1000).

At a test length of 80 items, Table 2 clearly shows that SEE Curves are highly stable. Similar to the effect noted with test lengths of 20, the expected decrease in variation of the standard errors with increase in sample size, is apparent only at ability levels of -1, +1, and +2.

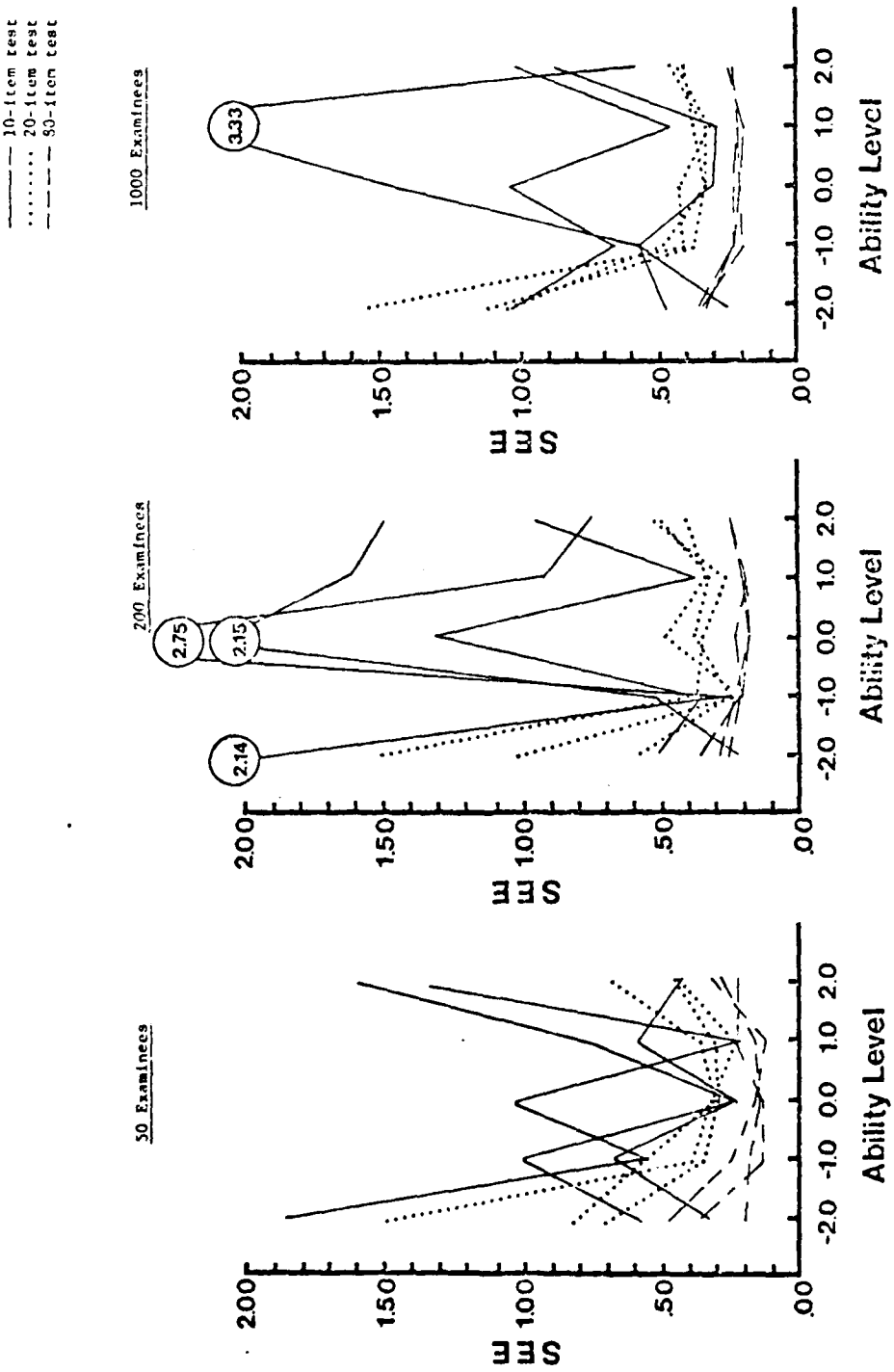

Figure 1. Standard errors of estimation associated with three test lengths (10, 20, and 80 test items) at five ability levels and reported for three sample sizes (50, 200, and 1000 examinees). (Each combination of condition was replicated three times.)

Table 7

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Homogeneous Item Pool (Test Length = 10 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	10	37	--	.66	.53	.51	1.43	1.05	2.71
	2	10	48	--	--	.79	.24	.48	1.59	--
	3	10	45	--	.80	.39	.42	.70	1.67	4.02
200	1	10	185	--	3.13	.41	.48	.41	1.16	4.03
	2	10	192	--	.52	.40	.65	.39	1.20	4.44
	3	10	179	--	3.89	.35	.60	.46	1.65	4.25
1000	1	10	960	--	--	.52	.46	.44	1.13	4.22
	2	10	960	--	--	.62	.41	.49	1.07	4.59
	3	10	996	--	--	.70	.40	.40	1.07	3.19

Table 8

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Homogeneous Item Pool (Test Length = 20 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	20	49	--	1.61	.45	.27	.44	.83	1.75
	2	20	49	--	.62	.56	.41	.33	.75	1.59
	3	20	50	--	2.53	.61	.17	.37	.63	1.52
200	1	20	194	--	1.83	.48	.34	.37	.70	1.60
	2	20	196	--	2.58	.47	.30	.39	.70	1.60
	3	20	198	--	2.64	.47	.30	.31	.69	2.03
1000	1	20	977	--	2.13	.46	.33	.33	.72	2.15
	2	20	984	--	2.09	.46	.34	.33	.68	2.01
	3	20	980	--	3.16	.53	.32	.33	.67	1.89

Table 9

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Homogeneous Item Pool (Test Length = 80 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	80	50	1.50	.69	.30	.16	.18	.31	.59
	2	80	50	.75	.31	.21	.18	.20	.40	.79
	3	80	50	1.14	.56	.26	.15	.22	.34	.64
200	1	80	200	1.17	.46	.23	.18	.21	.36	.68
	2	80	200	1.00	.40	.21	.20	.22	.37	.69
	3	80	200	1.08	.47	.24	.17	.20	.35	.72
1000	1	80	1000	1.21	.49	.23	.19	.20	.34	.71
	2	80	1000	1.24	.49	.23	.19	.20	.35	.71
	3	80	1000	1.13	.44	.23	.20	.21	.33	.69

Table 10

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Homogeneous Item Pool (Sample Size = 50 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	37	--	.66	.53	.51	1.43	1.05	2.71
	2	10	48	--	--	.79	.24	.48	1.59	--
	3	10	45	--	.80	.39	.42	.70	1.67	4.02
20	1	20	49	--	1.61	.45	.27	.44	.83	1.75
	2	20	49	--	.62	.56	.41	.33	.75	1.59
	3	20	50	--	2.53	.61	.17	.37	.63	1.52
80	1	80	50	1.50	.69	.30	.16	.18	.31	.59
	2	80	50	.75	.31	.21	.18	.20	.40	.79
	3	80	50	1.14	.56	.26	.16	.22	.34	.64

Table 11

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Homogeneous Item Pool (Sample Size = 200 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	185	--	3.13	.41	.48	.41	1.16	4.03
	2	10	192	--	.52	.40	.65	.39	1.20	4.44
	3	10	179	--	3.89	.35	.60	.46	1.65	4.25
20	1	20	194	--	1.83	.48	.34	.37	.70	1.60
	2	20	196	--	2.58	.47	.30	.39	.70	1.60
	3	20	198	--	2.64	.47	.30	.31	.69	2.03
80	1	80	200	1.17	.46	.23	.18	.21	.36	.68
	2	80	200	1.00	.40	.21	.20	.22	.37	.69
	3	80	200	1.08	.47	.24	.17	.20	.35	.72

Table 12

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Homogeneous Item Pool (Sample Size = 1000 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	960	--	--	.52	.46	.44	1.13	4.22
	2	10	960	--	--	.62	.41	.49	1.07	4.59
	3	10	996	--	--	.70	.40	.40	1.07	3.19
20	1	20	977	--	2.13	.46	.33	.33	.72	2.15
	2	20	984	--	2.09	.46	.34	.33	.68	2.01
	3	20	980	--	3.16	.53	.32	.33	.67	1.89
80	1	80	1000	1.21	.49	.23	.19	.20	.34	.71
	2	80	1000	1.24	.49	.23	.19	.20	.35	.71
	3	80	1000	1.13	.44	.23	.20	.21	.33	.69

Table 13

Variation of Standard Errors of Estimates at Several Ability Levels for Different Test Lengths and Examinee Sample Sizes (Heterogeneous Item Pool)

Test Length	Sample Size	Ability Level[1] -2.0	-1.0	0.0	1.0	2.0	Average Variation Across Ability Levels
10	50	.68	.20	.39	.23	.50	.40
	200	.85	.10	.60	.50	.31	.47
	1000	.32	.04	.47	1.40	.19	.60
20	50	.36	.10	.01	.05	.11	.16
	200	.38	.06	.06	.04	.06	.12
	1000	.22	.05	.04	.02	.02	.09
80	50	.11	.04	.01	.05	.03	.06
	200	.04	.02	.02	.01	.00	.02
	1000	.01	.01	.00	.00	.00	.00

[1]Each entry in this section was obtained by calculating the standard deviation of standard errors of estimates across three replications for a particular test length and sample size.

Table 14

Variation of Standard Errors of Estimates at Several Ability Levels for Different Test Lengths and Examinee Sample Sizes (Homogeneous Item Pool)

Test Length	Sample Size	Ability Level[1] -2.0	-1.0	0.0	1.0	2.0	Average Variation Across Ability Level
10[2]	50		.17	.11	.41	.28	.24
	200		.03	.07	.03	.22	.09
	1000		.07	.03	.04	.03	.04
20	50	.78	.07	.10	.05	.08	.22
	200	.37	.00	.02	.04	.00	.09
	1000	.50	.03	.01	.00	.02	.11
80	50	.16	.04	.01	.02	.04	.05
	200	.03	.01	.01	.01	.01	.01
	1000	.02	.00	.00	.00	.01	.01

[1]Each entry in this section was obtained by calculating the standard deviation of standard errors of estimates across three replications for a particular test length and sample size.

[2]Standard deviations were not calculated for this test length at ability level -2 because of extreme fluctuations in the data.

(b) Item Pool One—Effect of Test Length

Examination of the results reported in Table 4 indicate that, for samples of size 50, as test length increased, variation in the SEE Curves decreased at all ability levels.

Tables 5 and 6, which represent the results of the simulations for sample sizes of 200 and 1000, clearly show the following trends: (1) the most stable SEE Curves were obtained for the longest test length; and (2) for all ability levels, variation in the SEE Curves decreased as test length increased.

Table 13 presents a summary of the data found in Tables 1-6. Entries in this table are the standard deviations of the standard errors of estimate obtained across the three replications of the various studies. Standard deviations are reported for each test length-sample size combination across five ability levels. Also included in Table 13 is the average of the standard deviations across ability levels for each test length-sample size combination. It is this latter value that is the focus of the following discussion.

Several trends are apparent from examination of the average variation of standard errors: (1) the variation decreased as test length increased for all sample sizes, (2) when test length was fixed at 10 items, sample size had little or no effect on the stability of the SEE Curves, and (3) sample size, generally, had a noticeable effect on the stability of the SEE Curves.

Figure 1 contains three graphs illustrating the effect of test length and sample size on the stability of the SEE Curves at five ability levels. Each graph represents a plot of the values of the SEE Curves obtained when sample size was held constant and test length was varied.

It is clear, from examination of these graphs, that sample size has little effect on the stability of SEE Curves of short tests (n=10). The effect of sample size on the stability of the standard errors was most apparent for the intermediate length test (n=20). For a long test (n=80) sample size showed the most pronounced effect when there was an increase from 50 to 200 examinees. An effect was also noticed when sample size was increased from 200 to 1000 examinees, however, the improvements in precision were more modest in size.

(c) Item Pool Two—Effect of Sample Size

Table 7 presents the results of the simulations involving test lengths of 10 items. It should be noted that no values are reported for ability level -3 and also that the only complete set of values at ability level -2 are reported for a sample size of 200. Values obtained at these ability levels fluctuated greatly and so they are not reported (a similar explanation applies to other results not reported). In summary, there was a substantial improvement in the precision of SEE Curves for increasing sample sizes. In fact, the improvements in precision of SEE Curves due to sample size for test lengths of 20 and 80 items are also clear from a study of Tables 8 and 9.

(d) Item Pool Two—Effect of Test Length

The results of this investigation are reported in Tables 10-12. These results are very similar to those obtained for item pool one and therefore will not be discussed to any great extent. It is important to note that for all sample sizes and at all ability levels there appears to be fairly consistent tendency for the stability of the SEE Curves to increase as test length was increased.

Table 14 summarizes the results reported in Tables 7-12. Data are arranged in Table 14 in the same manner in which they were arranged in Table 13. Examination of the average variation across ability levels, indicated that for all test lengths, sample size has a noticeable effect on the stability of the SEE Curves. In comparison to the results reported in Table 13, the effect of test length on the average variation across ability levels is not so apparent. The reason for this is the smaller variation observed for short tests with this particular item pool.

Effects of Statistical Characteristics of an Item Pool on Precision of SEE Curves

A comparison of the results reported in Tables 13 and 14, indicated that for tests of 20 and 80 items, the variation in the SEE Curves, averaged across ability levels, is very similar for both item pools. For test lengths of 10, the situation is quite different. In order to make the average variations across ability levels at this test length comparable for both item pools, these values were recomputed for item pool two, excluding the values obtained for ability level of -2. The recomputed average variation values are .33, .38, and .52 for sample sizes of 50, 200 and 1000 respectively. It is clear that, for short tests, the homogeneous item pool (pool one) resulted in smaller average variations than did the heterogeneous item pool. A second point worth noting, is that the heterogeneous item pool (pool two) provided more stable Standard Errors at an ability of -2 for test lengths of 10 or 20 items than did the homogeneous item pool. For test lengths of 80, the results appear to be about the same for both item pools. It

should also be noted that the homogeneous item pool generally results in greater stability of Standard Errors for ability levels between +1 and -1 than did the heterogeneous item pool.

Relationship Between Test Length and SEE Curves in Two Typical Item Pools

Figure 2 contains two graphs, representing item pools one and two. These graphs show the relationship between test length and SEE Curves. Item parameters were used to derive the Curves rather than estimates of the item parameters. The trends in the results are generally what one would expect. The value of the figure is the information it provides to test developers who must determine a test length.

Test lengths of 10 and 20 items, drawn from the heterogeneous item pool (item pool one) do not show the expected U shaped pattern exhibited by the curves obtained for these test lengths when the simulation involved a homogeneous item pool. The "humping" effect noted at the center of the ability distribution is due to the particular sample of items chosen. There are a few less items selected with difficulty values close to zero. It is quite apparent that the heterogeneous item pool provided smaller standard errors of across a wider range of abilities than did the homogeneous item pool.

Further insight into the effect of the item pool on the size of the standard errors can be obtained by examination of the graphs presented in Figure 3. Each graph represents one of the three different test lengths that was studied. The relationship between test length and SEE between +3 and -3 is graphed for both item pools on the same axes to facilitate comparison of the effect of the item pools. The decrease in

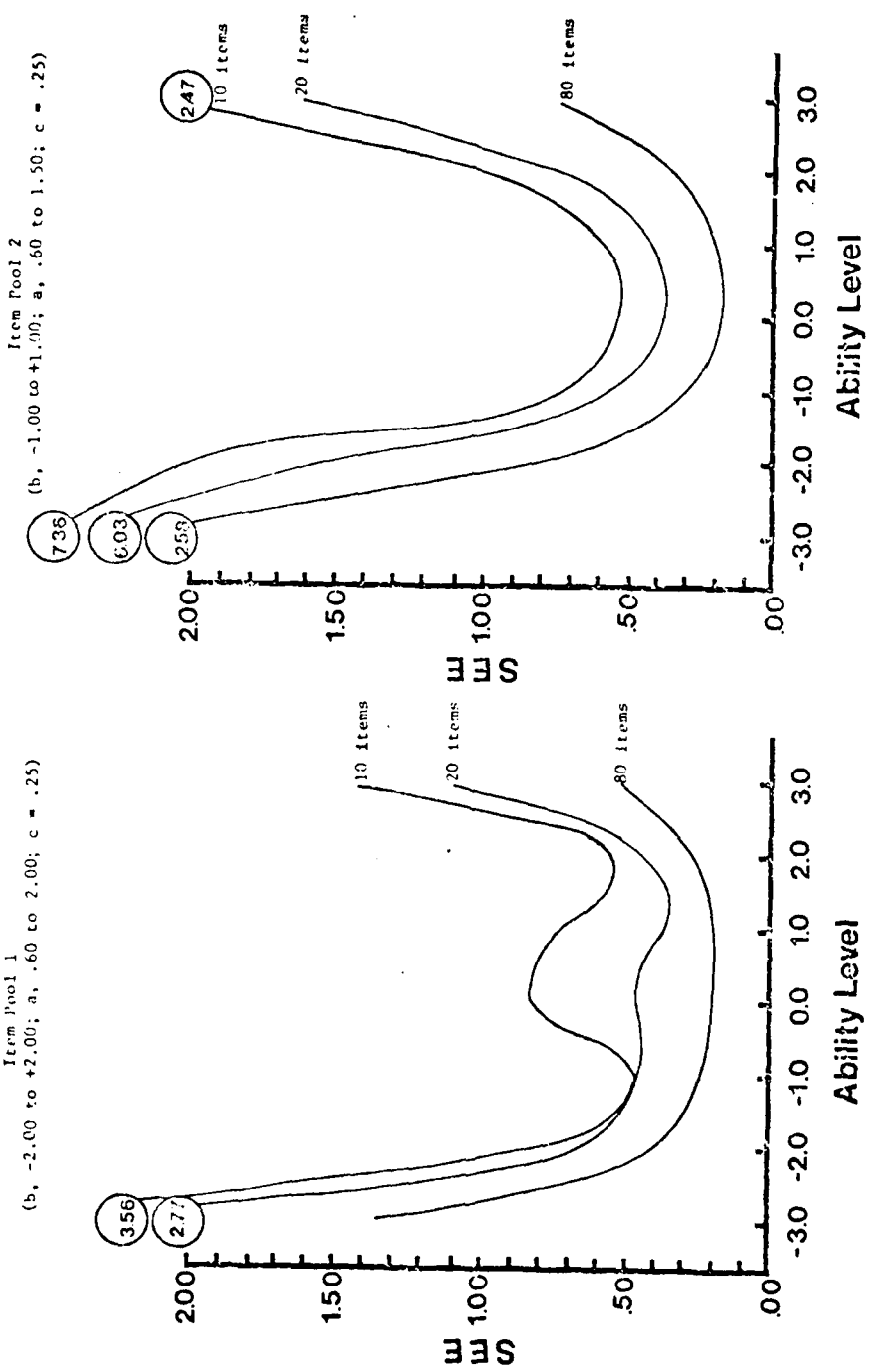

Figure 2. Standard errors of estimation associated with three test lengths at five ability levels and reported for two item pools.

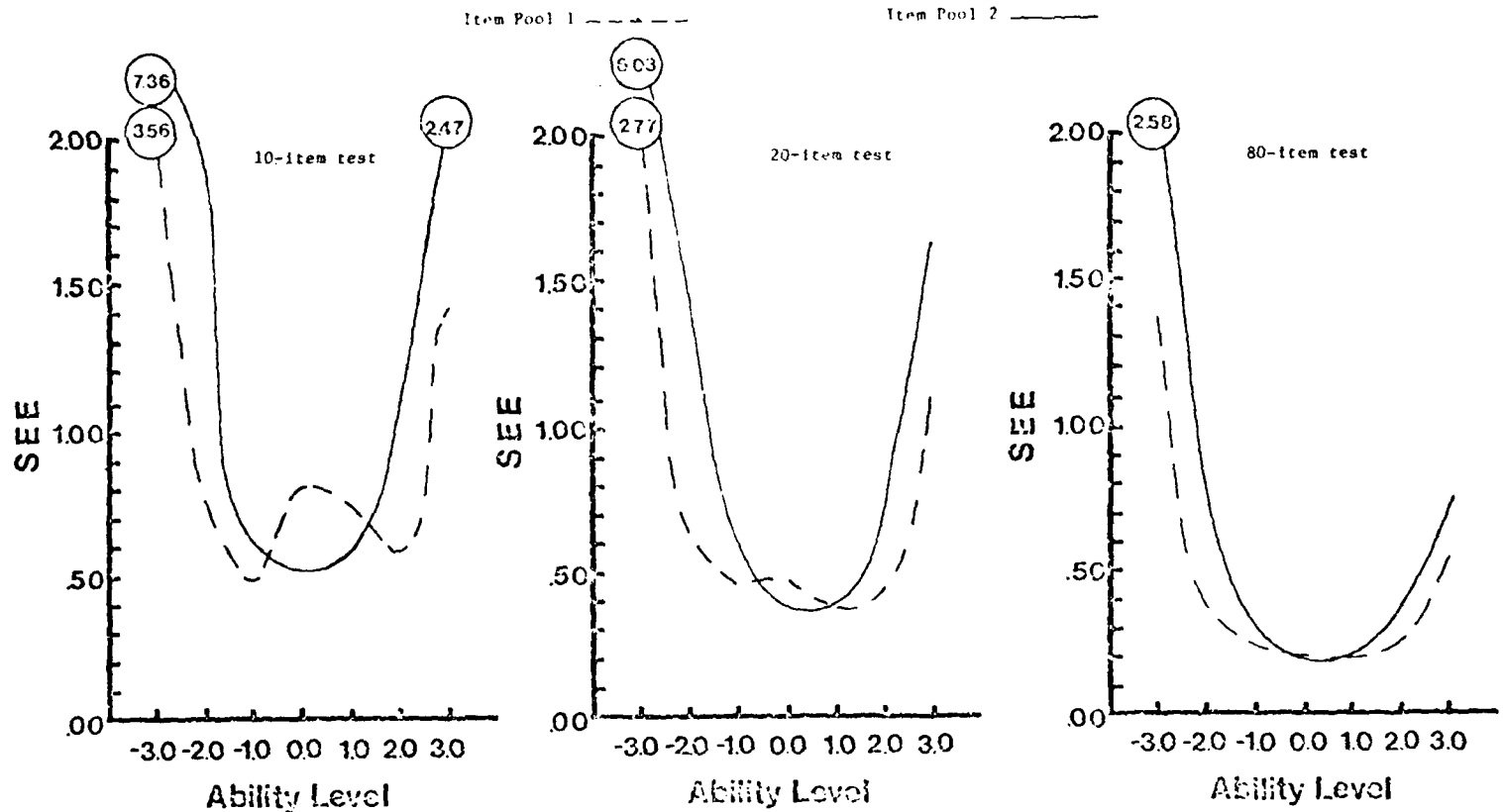

Figure 3. Standard errors of estimation associated with two item pools at five ability levels and reported for three test lengths.

the size of the standard errors as test length increases is quite evident for both pools. Also apparent is the fact that tests based on items drawn from the heterogeneous item pool provide greater precision over a wider ability range then do tests developed from the homogeneous item pool..

Conclusions

A study along the general lines as this one is not going to reveal any major new results. It is well-known that the size of an examinee sample, the length of a test, and the characteristics of an item pool, will have an important influence on the shape and stability of SEE Curves. The importance of this study is that it provides data concerning the size of improvements in SEE Curves relative to the three factors under investigation: (1) sample size, (2) test length, and (3) item pool characteristics. In this regard several conclusions seem warranted:

1. Both test length and sample size are extremely important factors in the precision of SEE Curves. (There were a small number of reversals in the results; no doubt this was due to sampling fluctuations.)

2. Precision of SEE Curves at the extremes of an ability continuum is very poor, even with large examinee sample sizes. The results are substantially better when tests are lengthened, even if the sample size is small (N=50).

3. The precision of SEE Curves would be acceptable in most instances if the Curves are based on 200 or more examinees with tests with at least 20 items. This recommendation holds if primary concern is with values of the Curves in middle regions of the ability continuum [-1 to +1].

4. Increases in examinee sample sizes from 50 to 200 produce sizeable improvements in the precision of SEE Curves. Gains in precision due to increasing a sample size from 200 to 1000 produce only modest gains in precision of the SEE Curves.

5. Similarly for test lengths, improvements in precision were substantially better when the change was from 10 to 20 items than 20 to 80 items.

Perhaps by offering a practical testing problem that arises, we can explain our interest in the precision of SEE Curves. Suppose a test developer selects a set of test items from a pool of items for a particular test he or she desires to build. Item selection is usually based on the item statistics. This test developer may then calculate the "expected" score information curve and corresponding SEE Curve. The usefulness of a SEE Curve will depend on its precision. If we knew that a second administration of the test to a similar group of examinees would produce a radically different curve, the curve will be of little or no value. The results of our study suggest that if an item pool is "typical," the stability of SEE Curves across readministrations of the test to similar groups of examinees will be quite good if the test includes at least 20 items, and if 200 or more examinees are used in deriving the item statistics.

We hope that our research has provided at least a few guidelines to aid test developers in determining the confidence which they should have in SEE Curves that arise in their work. If it also serves as a motivator to further extend our work by considering other aspects of the problem (for example, the shape of the underlying ability distribution, the number of parameters describing a test item, and methods used to estimate parameters) we will be even more pleased.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

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

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues, and applications. Review of Educational Research, 1979, in press.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, in press.

Marco, G. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Rentz, R. R., & Bashaw, W. L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work? A synthesis of the literature for practitioners. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement and Evaluation, Educational Testing Service, 1978.

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

Weiss, D. J. (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota, 1978.

Wright, B. D., & Stone, M. H. Best test design: A handbook for Rasch measurement. Palo Alto: Scientific Press, 1978.