The Effects of Item Calibration Sample Size and Item Pool Size on Adaptive Testing

Malcolm James Ree
Air Force Human Resources Laboratory

A simulation study of the effects of varying the item calibration sample size on varying size item pools was run for the maximum information adaptive test. Items were calibrated on the three-parameter logistic model on sample sizes of 500, 1,000, and 2,000. Item pools of 100, 200, or 300 items were developed from the three calibration sample sizes. Fixed-length adaptive tests of 10, 15, 20, 25, 30, and 35 items were given to a different group of 500 simulated subjects for each combination of item pool size and calibration sample size. Results indicated that high correlations between ability and estimated ability would be obtained in any testing if a sufficient number of items were administered. The reduction of absolute error of ability estimation was found to require at least 200 items calibrated on 2,000 subjects.

Adaptive, or tailored, testing refers to a series of techniques for finding and scoring the most useful set of items to be administered to an individual. Adaptive testing techniques range from strictly mechanical (Weiss, 1973) to mathematically elegant models (Owen, 1969). In all these techniques, the basic components are parameters representing the items, some method for ability estimation, a method for item selection, and rules for the initiation and termination of testing. Development of item pools can be very expensive, especially if large numbers of examinees are required in order to try out large numbers of experimental items. The object of this study is to investigate the consequences of item pool size and item calibration sample size on the outcomes of adaptive testing.

Frequently, items for adaptive testing are calibrated on a three-parameter model explained by Lord and Novick (1968) in the normal case and by Birnbaum (1968) in the logistic approximation. The three parameters of the Birnbaum model—*a*, *b*, and *c*—are item discrimination, item difficulty (or location), and probability of chance success (or lower asymptote), respectively. Equation 1 presents the mathematical function describing the item characteristic curve (ICC):

$$P(\theta) = c_i + (1 - c_i)(1 + e^{(-1.7a_i(\theta - b_i))})^{-1} \quad , \qquad [1]$$

where $P(\theta)$ is the probability of θ and the parameters a_i, b_i, and c_i refer to item discrimination, location, and lower asymptote, as previously noted, for item *i*. Previous research (Ree, 1979a; Urry, 1976) indicates that the ICC parameters may be estimated with some reasonable degree of accuracy, providing a sufficient sample of examinees with an appropriate distribution of ability (θ) is available.

For purposes of the present study, the maximum information adaptive test was selected,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 1, Winter 1981, pp. 11–19

based on previous research (Maurelli, 1978). The objectives of this study were to examine the consequences of differing calibration sample size and differing item pool size on the estimation of θ during adaptive testing. Simulation techniques were used in order to have known true or starting values for analysis and comparison.

Method

Item pool size was varied from 100 to 200 to 300 items and each of these pools was calibrated on 500, 1,000, and 2,000 examinees. A total of nine (three pool sizes × three calibration sample sizes) different item pools were available for the study.

Item Pool Calibration and Construction

The *a, b,* and *c* parameters for each of the items were generated as was a θ for each simulee. In order to generate a vector of item responses for each simulee, the θ values were used in Equation 1 to compute the likelihood of "correctly answering" each item. Because Equation 1 yields a number, $P(\theta)$, such that $0.0 < P(\theta) < 1.0$, a number, X, was drawn from a uniform (rectangular) distribution ranging from 0.0 to 1.0 and compared to $P(\theta)$. If X was larger than $P(\theta)$, then an incorrect response was specified for the item; otherwise, a correct response was specified for the item. Thus, a simulee with $P(\theta) = .90$ got the item correct 9 in 10 times, and a vector of item responses was developed for each simulee in each data set.

The procedure produced a vector of item responses (1 or 0) for every simulee. A total of 2,000 simulees were used in the three pools. The simulees were administered the items in three groups of 100 items each. This procedure follows the recommendations of the developer of the item calibration program OGIVIA (Urry, 1976) and the results of the previous investigations of the program (Ree & Jensen, 1979). The item sets were generated with the following specifications: the *a* parameters were to mirror what is found in current tests, the *b* parameters should be uniformly distributed over approximately a ±2 interval of θ, and the *c* parameter should be appropriate to a test item with five-item options.

Sample sizes of 500, 1,000, and 2,000 were randomly selected with replacement from each group of vectors. That is, each group of items was calibrated independently of the other two on the specified sample sizes. Based on past success with the OGIVIA program, the groups of items were calibrated and the χ^2 goodness-of-fit tests were examined for the exclusion of nonconforming items. No items had to be excluded even at the very extremes of the distribution or with the smaller sample sizes.

Three item pools of 100, 200, and 300 were then constructed by simply including the items in the groups of 100 in which they were generated. This procedure assured that the distribution of *b* parameters remained uniform regardless of the pool used. Item pools of these sizes were selected as appropriate for large-scale testing as might be required for military selection and classification.

In order to provide for comparisons with other studies of item parameter estimation (Ree & Jensen, 1979; Urry, 1976), two indexes of estimation success—the correlation between the estimated and true parameters and the Root Mean Squared Error (RMSE) of the estimate—as well as descriptive statistics, were computed.

The Maximum Information Adaptive Test and the Conventional Test

The adaptive test. The maximum likelihood procedure was similar to those proposed by Lord (1976) and by Samejima (1975). Item selection was based on choosing the item with the highest information value (Birnbaum, 1968) at the current estimate of θ, and ability was estimated by maximum likelihood.

All simulees entered the item pool with an initial estimate of θ at the mean of the ability distribution. Because the maximum likelihood ability estimation procedure requires at least one item answered correctly and one item answered incorrectly, the Bayesian ability estimation procedure

(Owen, 1969) was used until this condition was satisfied. Also, nonconvergences in the maximum likelihood estimation and extreme estimates in the Bayesian procedures were assigned either $\hat{\theta}$ values of +5.00 or −5.00. All tests were of fixed length, but because variable-length tests are of interest, six tests of different lengths were administered for each condition to each simulee. These lengths were 10, 15, 20, 25, 30, and 35 items, and the tests were all administered to 500 simulees randomly generated from the unit normal distribution ($\bar{\theta} = -.011, \sigma_\theta = .999$).

Two sets of item parameters were always used during each adaptive test administration. The true *a, b,* and *c* parameters were used to generate item responses in the same manner as referenced earlier, and estimated item parameters ($\hat{a}$, $\hat{b}$, $\hat{c}$) were used to estimate ability ($\hat{\theta}$) and to compute estimated item information ($\hat{I}$) for use in item selection.

Several indices were computed, including average estimated θ; the average algebraic difference between θ and $\hat{\theta}$, which is often called bias; the average absolute difference between θ and $\hat{\theta}$; and the correlation between θ and $\hat{\theta}$. These were denoted by $\bar{\hat{\theta}}$, $\overline{\theta - \hat{\theta}}$, $\overline{|\theta - \hat{\theta}|}$, and $r_{\theta \cdot \hat{\theta}}$, respectively.

The $\bar{\hat{\theta}}$ and $\overline{\theta - \hat{\theta}}$ were computed to investigate constant over- or underestimation, while $r_{\theta \cdot \theta}$ indicates, in general, how well $\hat{\theta}$ reproduces θ. Of all the indices computed, it is the most easily distorted by artifact. Simply raising or lowering the artificial nonconvergence values, which were set at ±5.0, alters the observed correlations. Because of this, the average absolute difference between θ and $\hat{\theta}$ was also computed. This index, $\overline{|\theta - \hat{\theta}|}$, provides a measure of the magnitude of the error that might be expected during test administration.

The conventional test. A conventional test (CT) of equal length to the adaptive tests was simulated for each simulee. Items were selected for the CT by starting with the item at the center of the pool (b = 0.0) and by moving up and down alternately to the next items. For example, a seven-item test would have items with the following b values: 0.0, .04, −.04, .08, −.08, .12, −.12. The longest CT had b values from −.70 to +.70 and varied *a* and *c* values representative of the item pool. Item responses were based on the true *a, b,* and *c* parameters, and CT scores were the sum of 1 or 0 item responses. No other ability estimates were made for the CT. The CT provided a base line for comparison of current practices with advanced adaptive testing techniques. For comparative purposes, the correlation between θ and the number-right (NR) score of the CT was computed. This value is not influenced by the potential artifactual distortion of $r_{\theta \cdot \hat{\theta}}$. The correlation between θ and the NR score did not make use of estimated item parameters and always used exactly the same items, whether in the 100, 200, or 300 item pool.

Results

Item Calibration

Table 1 provides the descriptive statistics of the true parameters, the parameters estimated on all samples, the intercorrelations between the true and estimated parameters, and root mean square error (RMSE) between true and estimated parameters. The results of the calibration were as expected and were consistent with past findings (Ree, 1979a; Ree & Jensen, 1979; Urry, 1976). Note that both the RMSE and correlational values were substantially the same for the three item pools, indicating, as do the parameter means and standard deviations across the item pools, the similarity of items in the three pools.

Maximum Information Adaptive Test

Tables 2, 3, and 4 present the results of the adaptive testing simulation. In all cases, correlations between θ and $\hat{\theta}$ increased (1) with increasing calibration sample size and (2) when a larger number of items was administered. The same may be observed in the measure of average absolute deviation of θ from $\hat{\theta}$.

Table 3 presents the correlation between θ and the NR score on the CT. Although they are presented in the table for the pool of 200 items in

Table 1
Descriptive Statistics of True and Estimated ICC Parameters, Their Intercorrelations, and RMSE

Parameter	Sample	Maximum	Minimum	Mean	Standard Deviation	r	RMSE
			Items 1 to 100				
a	True	2.000	.800	1.221	.338		
	2000	2.153	.714	1.306	.341	.890	.180
	1000	2.157	.710	1.259	.331	.856	.183
	500	2.058	.578	1.227	.324	.684	.262
b	True	1.960	-2.000	- .020	1.161		
	2000	2.280	-1.944	.072	1.195	.994	.160
	1000	2.161	-1.958	.052	1.217	.990	.202
	500	2.407	-1.954	.104	1.259	.987	.251
c	True	.293	.125	.195	.030		
	2000	.326	.059	.199	.062	.298	.060
	1000	.331	.046	.185	.080	.266	.078
	500	.344	.046	.185	.085	.188	.085
			Items 101 to 200				
a	True	2.000	.800	1.231	.344		
	2000	2.233	.720	1.308	.345	.882	.184
	1000	2.022	.691	1.261	.326	.873	.172
	500	2.121	.577	1.244	.333	.697	.263
b	True	1.960	-2.000	- .020	1.161		
	2000	2.299	-1.951	.064	1.200	.994	.155
	1000	2.185	-1.953	.058	1.218	.990	.202
	500	2.366	-1.950	.188	1.254	.987	.250
c	True	.297	.123	.195	.031		
	2000	.327	.056	.196	.064	.317	.061
	1000	.331	.046	.189	.078	.269	.076
	500	.346	.046	.188	.085	.195	.085
			Items 201-300				
a	True	2.000	.800	1.231	.346		
	2000	2.161	.690	1.309	.343	.921	.157
	1000	2.114	.670	1.277	.328	.882	.170
	500	2.951	.573	1.247	.357	.623	.304
b	True	1.960	-2.000	- .020	1.161		
	2000	2.332	-1.944	.074	1.193	.994	.159
	1000	2.335	-1.942	.093	1.224	.991	.206
	500	2.187	-1.961	.096	1.256	.987	.244
c	True	.280	.110	.193	.035		
	2000	.320	.060	.200	.063	.350	.060
	1000	.331	.052	.193	.076	.274	.074
	500	.342	.047	.181	.084	.184	.086

the rows for the calibration sample of 2,000 subjects, they would be equally appropriate placed alongside the other calibration sample sizes and in the other item-pool-size tables. This is because exactly the same items were administered each time a test of a particular length was re-

Table 2
Means, Deviation Measures, and Correlations
for a Pool of 100 Items

Number of	Maximum Information Adaptive Test			
Items	$\overline{\hat{\Theta}}$	$\overline{\lvert\Theta - \hat{\Theta}\rvert}$	$\overline{\Theta - \hat{\Theta}}$	$r_{\Theta\cdot\hat{\Theta}}$
2,000 subjects				
10	.127	.306	-.139	.936
15	.088	.263	-.099	.946
20	.077	.243	-.088	.957
25	.104	.220	-.115	.967
30	.067	.207	-.078	.969
35	.075	.201	-.086	.971
1,000 subjects				
10	.165	.334	-.176	.924
15	.131	.276	-.142	.952
20	.122	.253	-.133	.959
25	.131	.245	-.143	.964
30	.112	.213	-.124	.973
35	.102	.207	-.114	.976
500 subjects				
10	.180	.352	-.191	.927
15	.142	.287	-.154	.952
20	.149	.279	-.161	.959
25	.153	.258	-.164	.965
30	.124	.226	-.135	.972
35	.123	.230	-.134	.972

quired. In all cases, the correlations between θ and the scores on the CT were lower than the correlations between θ and $\hat{\theta}$ for equal length adaptive tests.

Discussion

Item Calibration

The indexes provided for comparison must be interpreted carefully. The correlation of a parameter and an estimate of that parameter would be misleading if a constant bias were evident. RMSE was used to obtain a measure of bias and variability; but as the two are confounded, care must be taken to interpret the correlation and RMSE together. These two indexes support the idea of reduction of error of estimation with increasing sample size.

It is interesting to note that there was an increasing mean difference (bias) between a and $\hat{a}$ with increasing sample size, accompanied by a reduction in RMSE with increasing sample size. The same decline in RMSE was found for the b and c parameters with increasing sample size and generally reduced mean difference in larger samples. The biased estimates of a did not seem to overly affect estimates of θ in longer adaptive tests. This is consistent with past results (Ree, 1979b) but requires more exhaustive research.

Table 3
Means, Deviation Measures, and Correlations
for a Pool of 200 Items

Number of	Maximum Information Adaptive Test				Conventional Test
Items	$\overline{\hat{\theta}}$	$\overline{\lvert\theta - \hat{\theta}\rvert}$	$\overline{\theta - \hat{\theta}}$	$r_{\theta \cdot \hat{\theta}}$	$r_{NR.\theta}$
2,000 subjects					
10	111	.282	-.123	.936	.855
15	.097	.224	-.108	.963	.879
20	.085	.199	-.097	.970	.906
25	.080	.182	-.091	.975	.913
30	.079	.169	-.091	.979	.921
35	.079	.166	-.090	.981	.930
1,000 subjects					
10	.130	.305	-.141	.923	
15	.121	.248	-.133	.956	
20	.125	.229	-.136	.968	
25	.116	.207	-.127	.973	
30	.113	.195	-.125	.978	
35	.116	.186	-.128	.981	
500 subjects					
10	.142	.300	-.153	.938	
15	.134	.260	-.145	.960	
20	.154	.245	-.166	.967	
25	.142	.243	-.154	.967	
30	.133	.222	-.145	.972	
35	.135	.217	-.147	.976	

Maximum Information Adaptive Test

The values in columns 2 ($\overline{\hat{\theta}}$) and 4 ($\overline{\theta - \hat{\theta}}$) in Tables 2, 3, and 4 indicate that maximum information adaptive testing procedures tend toward unbiasedness. Others (McBride, 1976; Maurelli, 1978) have found the same result with both similar and differing adaptive testing strategies. As more items were administered, the measure of bias, column 4, approached zero. The reduction was not greatly affected by increasing the item pool size. That may be because the smallest item pool was sufficiently large for most applications. However, it was affected by item calibration sample size and number of items administered. The lowest—hence, the least biased—values were found in the 300-item pool calibrated on 2,000 subjects.

The correlations of θ with $\hat{\theta}$, as shown in column 5 of Tables 2 through 4, show an increasing relationship with number of items in the pool, number of items administered, and number of subjects in the item calibration sample. The tabled values are all quite high, and, as expected, the highest values observed were in the 300-item pool when the largest number of items was administered. For comparative purposes, the correlation between CT scores and θ were uniformly lower for each number of items administered from 10 through 35.

The values in column 3 of Tables 2, 3, and 4 show the average absolute error of $\hat{\theta}$. This may

Table 4
Means, Deviation Measures, and Correlations
for a Pool of 300 Items

Number of	Maximum Information	Adaptive Test		
Items	$\hat{\theta}$	$\overline{\lvert\theta - \hat{\theta}\rvert}$	$\overline{\theta - \hat{\theta}}$	$r_{\theta \cdot \hat{\theta}}$
2,000 subjects				
10	.087	.294	-.098	.915
15	.096	.237	-.107	.953
20	.084	.220	-.095	.972
25	.077	.189	-.088	.973
30	.078	.179	-.091	.980
35	.080	.169	-.089	.982
1,000 subjects				
10	.126	.289	-.138	.908
15	.118	.250	-.129	.952
20	.121	.221	-.133	.970
25	.118	.207	-.129	.975
30	.116	.190	-.128	.980
35	.116	.185	-.128	.982
500 subjects				
10	.140	.302	-.158	.936
15	.145	.271	-.157	.923
20	.130	.240	-.141	.965
25	.133	.222	-.144	.971
30	.137	.210	-.148	.979
35	.136	.104	-.147	.981

be interpreted as the expected error involved in estimation of θ and is similar to the RMSE of θ reported by Urry (1976). These values indicate the magnitude of the average error that might be expected in estimation of θ during an adaptive testing session. This would be important for educational diagnosis or for personnel selection, classification, and placement. Greater errors in estimating ability would yield greater errors in classification. The magnitude of the errors was about as would be expected from past research, even though past research used known item parameters. McBride (1976) computed the index and found similar values but with a fully Bayesian adaptive test using a "perfect" item pool, item b parameters being defined as equal to the $\hat{\theta}$ previously computed. In order to determine if the current values were in error, the maximum likelihood program used in Koch and Reckase (1979), SIM3P, was obtained, the index ($\lvert\theta - \hat{\theta}\rvert$) incorporated, and several known, not estimated, data sets were run on both SIM3P and the author's procedure. SIM3P consistently gave errors of slightly larger magnitude. This might be attributed to the use of the different item selection technique used prior to making maximum likelihood ability estimates. SIM3P uses a fixed step size whereas the procedure used in this study (and Maurelli's, 1978, study) uses a Bayesian ability estimation technique until maximum likelihood estimates can be made. It was concluded that the magnitude of the errors was a function of the estimated item parameters, as intended in the study.

The values for average absolute error decreased with sample size, with number of items administered, and with addition of items to the pool, up to 200. Unexpectedly there was little or no decrease—and, in some instances, trivial increases—in this index when the pool was increased to 300 items. This is perhaps because adequate coverage of the ability continuum was achieved with 200 items; and adding items, in effect, added error. That is, sufficient items were always available to make good estimates with low error when 200 items were in the pool. Care must be taken not to generalize this to item pools of different quality and with different distributions of estimated item parameters. The effects of item pool construction have been only infrequently studied (Jensema, 1976), and not with the average absolute deviation index. However, larger item pools, in general, produce smaller error of ability estimation.

Clearly, the size of the item calibration sample as well as the number of items administered has been shown to be important to adaptive testing outcomes. Additional research with smaller item pools, ranging from about 25 to 100 (which might be constructed for classroom use), is needed. Also needed is research pertaining to the effects of imperfect estimation of each item parameter on the estimation of θ during adaptive testing.

If an ordering of examinees is all that is required or if the relatively higher errors are not important to the purpose, item pools of 100 items calibrated on a sample of 500 subjects will produce high correlations, especially if 20 or more items are administered. On the other hand, if high accuracy of point estimate of ability is required, then larger item calibration sample sizes and item pools are mandatory. It may be concluded that item pool size and item calibration sample size should be a function of the use to which the scores will be put.

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.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U.S. Civil Service Commission, Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Koch, W. R., & Reckase, M. D. *Problems in application of latent-trait models to tailored testing* (Research Report 79-1). Columbia: University of Missouri, Department of Psychology, 1979.

Lord, F. A broad range test of verbal ability. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U.S. Civil Service Commission, Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

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

McBride, J. R. Adaptive testing research at Minnesota—Some properties of a Bayesian sequential adaptive mental testing strategy. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U.S. Civil Service Commission, Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Maurelli, V. A. *A comparison of Bayesian and maximum likelihood scoring in a simulated stradaptive test.* Unpublished master's thesis, St. Mary's University of Texas, San Antonio, TX, 1978.

Owen, R. J. *A Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service, 1969.

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

Ree, M. *The effects of errors in estimation of item characteristic curve parameters.* Paper presented at the 21st annual convention of the Military Testing Association, San Diego, CA, October 1979. (b)

Ree, M. J., & Jensen, H. E. The effects of sample size on linear equating of item characteristic curves. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Computerized Adaptive Testing Laboratory, 1980. (Also published as AFHRL-TR-79-70, Brooks AFB, TX, 1979.)

Samejima, F. *Behavior of the maximum likelihood estimate in a simulated tailored testing situation.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, April 1975.

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U.S. Civil Service Commission, Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Weiss, D. J. *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

Acknowledgments

The opinions expressed are not necessarily those of the Department of the Air Force or of the United States Government. The author thanks Dr. C. J. Mullins, AFHRL; T. Warm, U.S. Coast Guard; and two unknown reviewers for their critical reading of earlier drafts of this manuscript.

Author's Address

Send requests for reprints or further information to Malcolm James Ree, Manpower and Personnel Division, Air Force Human Resources Laboratory, Brooks Air Force Base, TX 78235.