EFFECTIVENESS OF THE ANCILLARY ESTIMATION PROCEDURE [1]

JOHN F. GUGEL, FRANK L. SCHMIDT, AND VERN W. URRY
U.S. Civil Service Commission

Urry (1974a) has presented a graphic method to provide approximations for the item parameters of the normal ogive and Birnbaum logistic three-parameter latent trait models. This method has since been further developed (Urry, 1975) to provide a more accurate computational procedure for estimating the three parameters, a_i (item discriminatory power), b_i (item difficulty), and c_i (item coefficient of guessing). Programmed for the computer, this technique produces parameter estimates quickly and inexpensively.

Initial studies of this procedure employed large sample sizes (N=2000 and 3000 cases) and a relatively large number of items (n=100). Under these conditions, the procedure produces very accurate parameter estimates (Urry, 1975). We are now in a position to examine the effects of reduced numbers of cases and items on error in the parameter estimates and on the accuracy of tailored testing using those estimates. It is known *a priori*, of course, that reduction in either the number of cases or the number of items will, other things being constant, tend to increase estimation errors. But it is not known at present how large or practically significant such increases would be. The present study, exploratory in nature, is addressed to these questions.

METHOD

Based on suggestions by Lord (1968, p. 1016) and the results of the previous study by Urry (1975), it was decided to allow the number of items to vary from 50 to 100 and the number of cases to range from 500 to 2000. The initial 100-item bank, from which the smaller banks were later selected, was characterized by a_i values ranging uniformly from .80 to 2.20, b_i values distributed uniformly from −1.9 to +1.9, and c_i values from .02 to .24, also uniform in distribution. These parameter values are not different from what one might reasonably expect to find empirically given prescreening of items (Urry, 1974*a*; Jensema, 1972). In the reduced item samples, the a_i values were chosen in equal steps from .80 to 2.20. For example, there were five levels of a_i for the 50-item test and ten for the 100-item test. Ten values of b_i in equal steps between −1.9 and 1.9, inclusive, were arranged within each level of a_i. (an exception was the 55-item test, which had eleven values of b_i in equal steps between −1.9 and 1.9, inclusive, within each of its a_i values.) For different levels of a_i, items were matched on b_i values. The c_i values ranged from .02 to .24 in equal steps, irrespective of a_i and b_i. Values of θ, representing simulated subjects, were sampled randomly from $N(0,1)$. Then for each θ, the simulation procedure described by Urry (1975 was used to generate a vector of responses (1 = correct; 0 = incorrect) for the item bank in question using the known item parameters. Parameter estimation was then carried out using this simulated data.

Two indices were used to evaluate the parameter estimates relative to the known parameters. First, the root mean square error (RMSE) was computed for the estimated parameters. The formula for this statistic, is;

$$\text{RMSE} = \sum_{1}^{n}\left(\frac{(p-\hat{p})^2}{n}\right)^{1/2} \qquad (1)$$

where the p = known values of a_i, b_i, c_i, or $\rho_{I\theta}$, and
n = number of items involved in the particular analyses.

Second, Pearson correlations between the known and estimated parameters were computed, i.e., $r_{p\hat{p}}$.

To illustrate the effects of error in the parameter estimates on the accuracy of tailored testing, Owen's (1968) algorithm was employed. Specifically, tailored testing was carried out on 100 simulated subjects using first the known item parameters and then item parameter estimates obtained on 1000 cases and 60 items. To increase the number of items used in tailored testing to a more realistic level, another identical set of 60 items was parameterized on a separate, independent group of 1000 simulated subjects, and these "items" were combined with the original 60 to produce a bank with 120 items. In the case of the known parameters, both 60-item sets were entered into the tailored testing bank. The known parameters in this bank were used to generate the response vectors of the 100 simulated subjects, and these vectors in turn, were used in the tailored testing. Correlations between estimated and actual θ were computed at each of eight termination rules for each condition of testing. This allowed a comparison of correlations across the conditions of testing, i.e., where (1) known or (2) estimated item parameters were used in the tailoring process.

[1] Computer processing for this study was done at the University of Maryland Computer Science Center in conjunction with graduate work by John Gugel. Arrangements for computer time were made by Professor Charles Johnson of the Department of Measurement and Statistics, College of Education, University of Maryland.

RESULTS AND DISCUSSION

Results produced by the parameterization procedure for varying combinations of sample size and number of items are shown in Tables 1 and 2. In both tables, "Raw Score Estimates" refer to the parameter estimates prior to application of the ancillary correction procedure, and the columns headed "Final Estimates" refer to estimates after application of the corrections. Table 1 includes the S.E. for $\rho_{I\theta}$, the correlation between the continuum underlying the item and θ, as well as for a_i, b_i, and c_i. "Lost items" are those for which the estimation procedure did not converge because of insufficient cases in the tails of the distribution.

Looking at the S.E.'s for the final estimates in Table 1, it can be seen that, in general, decreasing both sample size and number of items results in increased RMSE's. This effect appears to be more pronounced for a_i than for the other parameters. Moving from 50 to 60 items (sample size constant) appears to produce marked reductions in error for a_i, but beyond this, improvements in accuracy with increases in number of items are smaller. The b_i and c_i were estimated rather accurately throughout the range of both independent variables, although variation in sample size and number of items did have the expected effect. The last column in Table 1 reveals a tendency for items to begin to fail to converge during parameter estimation when sample size is dropped as low as 500. Sample size appears more crucial in this respect than number of items. Correlations between final parameter estimates and actual parameters, shown in Table 2, also pattern themselves as expected, within the limits of sampling error. In examining these correlations, one should bear in mind that in the case of $\hat{a}_i$ and to a lesser extent $\hat{c}_i$, restriction in range is operating to lower the tabled values. The items parameterized contained no values of a_i lower than .80. This value of a_i corresponds to a biserial correlation of .62 between the item and latent ability. Past studies (Jensema, 1972; Urry, 1974*b*) have shown that only about one third of the items in conventional tests have a_i values this large. No c_i greater than .24 were included; in practice c_i does exceed .24, although the range restriction here is probably less severe than in the case of a_i.

Results of simulated tailored testing using known parameters and parameters estimated on a sample of 1000 with 60 items are shown in Table 3. The eight termination rules, expressed as the standard error of estimate ($\sigma_{\hat{\epsilon}}$) are seen in column 2. Column 3 translates these values to reliability coefficients for $\hat{\theta}$, based on the relationship

$$\rho^2_{\hat{\theta}\theta} = 1 - \sigma_{\hat{\epsilon}}{}^2 \qquad (2)$$

TABLE 1

Root Mean Square Errors (RMSE)
Before and After all Corrections

		Raw Score Estimates RMSE				Final Estimates RMSE				
Items	Cases	a_i	b_i	c_i	$\rho_{I\theta}$	a_i	b_i	c_i	$\rho_{I\theta}$	Lost Items
50	2000	.283	.124	.086	.043	.395	.137	.064	.053	0
50	1000	.292	.193	.097	.053	.326	.209	.078	.059	1
50	500	.370	.164	.097	.067	.472	.259	.077	.064	0
55	2000	.385	.195	.091	.061	.308	.150	.057	.053	0
55	1000	.352	.194	.101	.050	.315	.124	.071	.050	0
55	500	.281	.185	.098	.054	.403	.227	.086	.065	4
60	2000	.321	.204	.091	.056	.253	.140	.065	.040	0
60	1000	.343	.231	.089	.059	.322	.144	.062	.044	0
60	500	.360	.194	.080	.070	.342	.179	.068	.062	0
70	2000	.272	.131	.095	.041	.225	.166	.067	.040	1
70	1000	.324	.189	.095	.054	.273	.174	.074	.045	1
70	500	.386	.197	.096	.072	.351	.187	.083	.058	4
80	2000	.266	.141	.092	.046	.214	.150	.072	.039	1
80	1000	.259	.178	.092	.048	.261	.166	.073	.047	1
80	500	.319	.224	.091	.063	.311	.229	.079	.048	6
90	2000	.297	.180	.094	.049	.244	.149	.069	.036	0
90	1000	.341	.171	.089	.051	.304	.140	.072	.044	0
90	500	.316	.184	.094	.056	.283	.144	.086	.049	2
100	2000	.290	.138	.085	.049	.223	.131	.056	.036	0
100	1000	.286	.137	.088	.052	.240	.162	.062	.039	0
100	500	.354	.189	.100	.061	.276	.161	.083	.047	5

TABLE 2

Correlations–Known Parameters vs. Estimated Parameters
Before and After All Corrections

Items	Cases	Raw Score Estimates			Final Estimates		
		$r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{c\hat{c}}$	$r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{c\hat{c}}$
50	2000	.846	.999	.599	.849	.997	.636
50	1000	.888	.992	.429	.908	.990	.492
50	500	.745	.993	.428	.780	.989	.454
55	2000	.731	.995	.488	.891	.995	.646
55	1000	.758	.995	.428	.870	.995	.546
55	500	.850	.992	.387	.824	.990	.376
60	2000	.828	.996	.491	.899	.997	.630
60	1000	.771	.994	.546	.842	.995	.588
60	500	.768	.994	.626	.801	.995	.668
70	2000	.834	.997	.471	.922	.997	.632
70	1000	.813	.996	.468	.828	.996	.521
70	500	.715	.993	.464	.813	.995	.449
80	2000	.873	.996	.535	.914	.997	.574
80	1000	.850	.994	.465	.879	.993	.550
80	500	.839	.991	.410	.823	.989	.502
90	2000	.861	.996	.483	.871	.996	.568
90	1000	.757	.995	.518	.847	.995	.547
90	500	.804	.995	.447	.874	.993	.418
100	2000	.837	.997	.539	.877	.998	.690
100	1000	.843	.996	.470	.863	.996	.627
100	500	.741	.993	.344	.824	.994	.420

The square root of this value is $\rho_{\hat{\theta}\theta}$, the correlation between the latent ability estimates ($\hat{\theta}$) and actual latent ability (θ). Validity coefficients of this sort are given in columns 4, 5, and 7. Those in column 4 are theoretical validities based solely on the termination rule chosen. Those in column 5 were obtained by correlating the $\hat{\theta}$ produced using the known item parameters with known θ. As expected they are essentially identical to the predicted theoretical validities. Those in column 7 were obtained by correlating the $\hat{\theta}$ produced using the parameter estimates with the known θ. As expected, they are somewhat lower than those in columns 4 and 5, but it can be noted that, as

TABLE 3

Validity Coefficients ($r_{\hat{\theta}\theta}$), and Average Number of Items ($\bar{n}$) Required for
Tailored Testing to Various Termination Rules Where the Item
Parameters Were Known or Estimated

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Termination Rules			Parameters Known		Parameters Estimated	
#	σ_ϵ	$\rho^2_{\hat{\theta}\theta}$	$\rho_{\hat{\theta}\theta}$	$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$
1	.5477	.70	.84	.864	2.43	.792	2.26
2	.5000	.75	.87	.904	3.31	.821	2.89
3	.4472	.80	.89	.932	4.00	.821	2.89
4	.3873	.85	.92	.935	4.91	.864	3.70
5	.3162	.90	.95	.955	7.03	.895	5.30
6	.2828	.92	.96	.962	8.77	.921	6.57
7	.2449	.94	.97	.969	11.77	.942	8.91
8	.2236	.95	.97	.975	14.51	.952	11.12

the termination rule becomes more stringent, the discrepancy decreases. At the most stringent termination rule, the validity of the $\hat{\theta}$ derived using the parameter estimates is only .023 lower than that based on the known parameters. The reliabilities of the two $\hat{\theta}$'s at this termination rule are .95 and .91, respectively.

Why are the termination rules not fully attained when the parameter estimates are used? The tailoring algorithm capitalizes on errors in the parameter estimates. As a consequence, tailored testing using the estimated parameters terminates prior to actually reaching the pre-set termination rule. That is, because of capitalization on error in parameter estimates during the process of item selection, the reliability levels implied by the Owen algorithm at any stage during the tailoring process are somewhat inflated. This leads to a too early termination of tailored testing, and, when the obtained $\hat{\theta}$ are correlated with θ, it becomes evident that the pre-set reliability level for termination has not been met. In the present example, an average of 14.51 items was administered when the known parameters were used but only 11.12 when the parameter estimates were used. This shrinkage problem can be overcome by setting the reliability termination rule higher than that actually required. In our present example, the termination rule should be set at .95 in order to obtain $\hat{\theta}$ of reliability .91.

REFERENCES

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished Doctoral Dissertation. University of Washington, 1972.

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.

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

Urry, V. W. *Computer assisted testing: the calibration and evaluation of the verbal ability bank.* (TS-74-3) Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974. (*b*)

Urry, V. W. Ancillary estimators for the item parameters of mental test models. Manuscript submitted for publication, 1975.