ITEM PARAMETERIZATION PROCEDURES FOR THE FUTURE

FRANK L. SCHMIDT AND VERN W. URRY
U.S. Civil Service Commission

Failure to appreciate the important psychometric role played by guessing in conventional multiple choice tests prevented until recently practical application of latent trait theory to tailored testing. When this problem was properly addressed, it was found that the solution could be expanded to produce an inexpensive and highly accurate item parameterization procedure. Combined with Owen's (1969) elegant Bayesian algorithm and available CRT hardware, these developments made computer-assisted tailored testing feasible from a practical point of view.

The capacity to parameterize new items for possible later inclusion in the item bank during routine operation of the computer-assisted testing system would be a significant step in the direction of even greater practicality (Killcross, 1974). Such a procedure would eliminate the necessity for periodic application of the full parameterization process described by Urry (1975*a*; 1975*b*). The Urry ancillary estimation procedure can be modified to provide the capability to parameterize items in the environment of a live, large-scale, computer-interactive tailored testing system or network. It can thus provide a convenient technology for updating and expanding item banks in ongoing tailored testing systems.

The parameterization procedure is as follows: In addition to the items that are part of his tailored test, each examinee receives a group of additional experimental items. On-line ancillary parameterization can begin for any of these items as soon as a sufficient number of examinees have responded to it. For each item, $\hat{\rho}_{i\theta}$ is computed against the uniformly reliable Bayesian θ from the Owen algorithm. (Notice that the item does not enter in any way into the determination of $\hat{\theta}$.) P_i' is estimated in the usual way using sample data. The $\hat{\theta}$ are next grouped into k intervals. Provisional values for $\hat{c}_i$ are assumed, and the minimum χ^2 procedure is applied to obtain approximations of a_i, b_i and c_i. These procedures have been outlined in Urry (1975*b*) and are described in full in Urry (1975*a*).

The purpose of this study was to evaluate the on-line ancillary parameterization process using model sampling and simulation techniques. The one hundred items to be parameterized were those used in the earlier Gugel study, and are shown in Table 1. (In practice, a much smaller number of items would typically be parameterized, but for evaluation purposes a larger number is desirable.) Dependent variables in this study were also the same as those in Gugel's study: correlations between known and estimated parameters and the square root of mean squared deviations of estimated from known parameters. Independent variables are illustrated in Figure 1. Two different banks were used in tailored testing to produce the Owen $\hat{\theta}$, designated as the Verbal Ability Bank and the Ideal Bank. The Verbal Ability Bank of this study consists of the 103 most frequently used items (based on counts from previous simulation studies) from the Commission's 200-item Verbal Ability Bank. The Commission's bank in turn, is made of the best 200 items out of 700 verbal ability items calibrated by Urry (1974). Calibration was carried out on large samples and the final 200 items were chosen to provide a wide distribution of b_i values, high a_i values, and low (below .30) c_i values. The 103 item bank used here thus represents a currently attained—though improvable—level of quality. The Ideal Bank is the same 100 items being parameterized (See Table 1). Three different termination rules were examined for the Ideal Bank; for the Verbal Ability Bank, the most stringent rule (.95) was omitted as impractical. Sample sizes of 1000, 1500, and 2000 were examined. Simulated subjects (θ's) were sampled and their response vectors generated as in the Gugel study. (This procedure is described in full in Urry [1974*a*]).

RESULTS AND DISCUSSION

The obtained standard errors for the Ideal and Verbal Ability banks are shown in Tables 2 and 4, respectively. Tables 3 and 5 present the correlations between actual and estimated item parameters. In most cases, changes associated with variation in the independent variables were in the hypothesized direction. Increasing the number of subjects and the reliabilities required for termination of tailored testing usually resulted in lower standard errors and higher correlations between known and estimated parameters. Some deviation from this pattern occurred because of sampling error. (For each bank, a different sample of simulated subjects was used for each termination rule and sample size examined.) The same is true of the ancillary corrections: the effect was generally to decrease standard errors and increase correlations, but because of sampling error this was not always the case.

In examining the correlations between known and estimated parameters, 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 (See Table 2) contained no values of a_i lower than .80. This value of a_i corresponds to a biserial

TABLE 1

True Parameters of the 100 Items Parameterized
Via the On-Line Procedure

Item	Parameters			Item	Parameters		
(i)	a_i	b_i	c_i	*(i)*	a_i	b_i	c_i
1	.80	-1.90	.03	51	1.60	.10	.18
2	.80	-1.70	.06	52	1.60	.30	.21
3	.80	-1.50	.09	53	1.60	.50	.24
4	.80	-1.30	.12	54	1.60	.70	.27
5	.80	-1.10	.15	55	1.60	.90	.03
6	.80	-.90	.18	56	1.60	1.10	.06
7	.80	-.70	.21	57	1.60	1.30	.09
8	.80	-.50	.24	58	1.60	1.50	.12
9	.80	-.30	.27	59	1.60	1.70	.15
10	.80	-.10	.03	60	1.60	1.90	.18
11	.80	.10	.06	61	2.00	-1.90	.21
12	.80	.30	.09	62	2.00	-1.70	.24
13	.80	.50	.12	63	2.00	-1.50	.27
14	.80	.70	.15	64	2.00	-1.30	.03
15	.80	.90	.18	65	2.00	-1.10	.06
16	.80	1.10	.21	66	2.00	-.90	.09
17	.80	1.30	.24	67	2.00	-.70	.12
18	.80	1.50	.27	68	2.00	-.50	.15
19	.80	1.70	.03	69	2.00	-.30	.18
20	.80	1.90	.06	70	2.00	-.10	.21
21	1.20	-1.90	.09	71	2.00	.10	.24
22	1.20	-1.70	.12	72	2.00	.30	.27
23	1.20	-1.50	.15	73	2.00	.50	.03
24	1.20	-1.30	.18	74	2.00	.70	.06
25	1.20	-1.10	.21	75	2.00	.90	.09
26	1.20	-.90	.24	76	2.00	1.10	.12
27	1.20	-.70	.27	77	2.00	1.30	.15
28	1.20	-.50	.03	78	2.00	1.50	.18
29	1.20	-.30	.06	79	2.00	1.70	.21
30	1.20	-.10	.09	80	2.00	1.90	.24
31	1.20	.10	.12	81	2.40	-1.90	.27
32	1.20	.30	.15	82	2.40	-1.70	.03
33	1.20	.50	.18	83	2.40	-1.50	.06
34	1.20	.70	.21	84	2.40	-1.30	.09
35	1.20	.90	.24	85	2.40	-1.10	.12
36	1.20	1.10	.27	86	2.40	-.90	.15
37	1.20	1.30	.28	87	2.40	-.70	.18
38	1.20	1.50	.06	88	2.40	-.50	.21
39	1.20	1.70	.09	89	2.40	-.30	.24
40	1.20	1.90	.12	90	2.40	-.10	.27
41	1.60	-1.90	.15	91	2.40	.10	.03
42	1.60	-1.70	.18	92	2.40	.30	.06
43	1.60	-1.50	.21	93	2.40	.50	.09
44	1.60	-1.30	.24	94	2.40	.70	.12
45	1.60	-1.10	.27	95	2.40	.90	.15
46	1.60	-.90	.03	96	2.40	1.10	.18
47	1.60	-.70	.06	97	2.40	1.30	.21
48	1.60	-.50	.09	98	2.40	1.50	.24
49	1.60	-.30	.12	99	2.40	1.70	.27
50	1.60	-.10	.15	100	2.40	1.90	.03

correlation of .62 between the item and latent ability. Past studies (Jensema, 1972; Urry, 1974) have shown that only about one-third of the items in conventional tests have a_i values this large. No c_i greater than .27 were included; in practice c_i does exceed .27, although the range restriction here is probably not as great as in the case of a_i.

The rather high a_i values among the items parameterized must be considered also in evaluating the root mean square

		Cut-offs*	ITEM BANKS	
			IDEAL BANK	VERBAL ABILITY BANK
S		.91		
	1000	.93		
U		.95		
B		.91		
J	1500	.93		
		.95		
E		.91		
C	2000	.93		
T		.95		
S				

*Reliability values for termination rules.

Figure 1. Experimental Design: Independent Variables

errors for a_i. Errors in $\hat{a}_i$ are much larger for high a_i than low a_i, since when a_i is high, small errors in $\hat{\rho}_{I\theta}$ lead to large errors in $\hat{a}_i$. For example, if $\hat{\rho}_{I\theta} = .90$, $a_i = 2.01$. If $\hat{\rho}_{I\theta} = .88$, $\hat{a}_i = 1.85$, a difference of .16. But if $\hat{\rho}_{I\theta} = .50$, $a_i = .58$. Then if $\hat{\rho}_{I\theta} = .48$, $\hat{a}_i = .55$, a difference of only .03.

The real test of the usefulness of the on-line parameterization process lies in the performance of the parameter estimates in tailored testing. The better the estimates, the closer they will come to equaling the performance of the known parameters. The parameter estimates obtained in this study have not yet been used in simulated tailored testing, but an idea of how well they would perform can be obtained by examining the performance of parameter estimates from Gugel et al. (1975) with roughly equivalent errors. Table 6 compares root mean square errors and correlations between known and estimated parameters from the present study for the Verbal Ability Bank with 2000 cases and reliability cut-off of .93 with the results obtained by Gugel et al. (1975) using 1000 cases and 60 items with the full parameterization process. Except for the standard error of $\hat{b}$ (which is lower) and $r_{\hat{a}a}$ (which is also lower), his results are essentially equivalent. Using a reliability cut-off of .95, Gugel et al. conducted simulated tailored testing using both the known and the estimated parameters. Known parameters produced $r_{\hat{\theta}\theta} = .9752$, exactly corresponding to the termination rule (i.e., $[.9752]^2 = .95$). With the parameter estimates, $r_{\hat{\theta}\theta}$ was .9516, corresponding to an obtained reliability of .9044.

Because the tailoring algorithm capitalizes on chance errors in the parameter estimates, tailored testing using the estimated parameters is terminated 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 computed 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 14.57 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 .90. Simulation studies provide a convenient—and perhaps the only—method of determining in advance of actual use the amount of shrinkage to be expected when items are parameterized on given sample sizes and with given numbers of items. The shrinkage problem here is thus somewhat different from that characterizing, say, multiple regression, in that its effects can be cancelled out by appropriate selection of termination rules. Two points, however, should be noted here:

1. Parameterizing on large sample sizes (both numbers of items and numbers of cases), and thus obtaining more accurate initial parameter estimates, is preferable where feasible to adjusting termination rules to allow for shrinkage.

2. For certain tailored testing usages—for example, battery tailoring or multivariate tailored testing—the advantages of parameter estimates that can fully meet pre-set termination rules become substantial. That is, adjustment of termination rules to allow for shrinkage becomes, at best, inconvenient and awkward.

In light of these facts, an important question is whether or not the on-line parameterization process can produce estimates with errors low enough to reduce shrinkage to negligible levels. An important consideration, of course, is the quality of the item bank on which the original $\hat{\theta}$ are derived. By parameterizing and adding to the Verbal Ability Bank those items which were erroneously rejected earlier on the basis of low point-biserial and biserial item-total indices, it will probably be possible to make the Verbal Ability Bank equivalent to the Ideal Bank used in this study. By increasing the number of cases to 3000, or perhaps beyond 3000, it should be possible to reduce the

TABLE 2

Root Mean Square* For Item Parameter Estimates And $\hat{\rho}_{I\theta}$ Using the Ideal Bank

Subject	Cut-offs	Uncorrected a_i	Uncorrected b_i	Uncorrected c_i	Uncorrected $\rho_{I\theta}$	Corrected a_i	Corrected b_i	Corrected c_i	Corrected $\rho_{I\theta}$
	.91	.465	.226	.089	.076	.340	.174	.086	.057
1000	.93	.480	.227	.095	.075	.357	.164	.079	.054
	.95	.418	.202	.093	.068	.283	.187	.074	.045
	.91	.481	.189	.086	.079	.318	.225	.075	.051
1500	.93	.467	.202	.091	.079	.290	.208	.067	.049
	.95	.445	.193	.095	.071	.311	.206	.070	.047
	.91	.506	.232	.091	.082	.267	.236	.079	.044
2000	.93	.477	.218	.090	.071	.270	.198	.067	.042
	.95	.454	.209	.090	.071	.297	.203	.066	.042

*RMSE = $\left(\frac{\Sigma(p_i-\hat{p}_i)^2}{n}\right)^{1/2}$

where p = parameters
n = number of items

TABLE 3

Correlations Between Known and Estimated Parameters–Ideal Bank

Subject	Cut-offs	*Uncorrected* a_i	*Uncorrected* b_i	*Uncorrected* c_i	Corrected a_i	Corrected b_i	Corrected c_i
	.91	.807	.995	.567	.820	.994	.548
1000	.93	.780	.994	.495	.780	.994	.540
	.95	.876	.994	.504	.874	.995	.553
	.91	.844	.996	.617	.832	.995	.656
1500	.93	.861	.995	.593	.860	.995	.624
	.95	.857	.995	.567	.852	.995	.610
	.91	.883	.995	.610	.886	.995	.631
2000	.93	.892	.996	.602	.892	.996	.641
	.95	.883	.996	.617	.883	.997	.649

TABLE 4

Root Mean Square Errors* For Item Parameter Estimates And $\hat{\rho}_{I\theta}$ Using the Verbal Ability Bank

Subject	Cut-offs	Uncorrected a_i	Uncorrected b_i	Uncorrected c_i	Uncorrected $\rho_{I\theta}$	Corrected a_i	Corrected b_i	Corrected c_i	Corrected $\rho_{I\theta}$
	.91	.596	.259	.093	.103	.370	.261	.097	.055
1000	.93	.599	.285	.093	.107	.400	.258	.095	.060
	.91	.514	.208	.090	.081	.280	.267	.084	.048
1500	.93	.554	.286	.082	.098	.336	.267	.075	.050
	.91	.562	.217	.087	.096	.338	.275	.076	.043
2000	.93	.553	.257	.086	.096	.331	.250	.072	.045

*RMSE = $\left(\frac{\Sigma(p_i-\hat{p}_i)^2}{n}\right)^{1/2}$

where p = parameter,
n = number of items.

TABLE 5

Correlations Between Known And Estimated Parameters–Verbal Ability Bank

Subject	Cut-offs	Uncorrected a_i	Uncorrected b_i	Uncorrected c_i	Corrected a_i	Corrected b_i	Corrected c_i
	.91	.786	.993	.524	.780	.933	.550
1000	.93	.821	.993	.510	.807	.993	.515
	.91	.875	.994	.565	.875	.994	.594
1500	.93	.871	.993	.614	.870	.993	.624
	.91	.836	.996	.622	.819	.995	.655
2000	.93	.878	.996	.562	.879	.996	.591

TABLE 6

Comparison of Gugel Results with Present Study Results

	Root Mean Square Errors a_i	b_i	c_i	$\rho_{I\theta}$	Correlations ($r_{\hat{p}p}$) $r_{\hat{a}_i a_i}$	$r_{\hat{b}_i b_i}$	$r_{\hat{c}_i c_i}$
Gugel (1975)*	.322	.140	.062	.044	.842	.995	.588
Present Study**	.331	.250	.072	.045	.879	.996	.591

*N = 1000, 60 items; full parameterization procedure.
**Verbal Ability Bank, N = 2000, Relaibility cut-off = .93.

root mean square errors shown in Table 2 (2000 cases, cut off at .95) to levels comparable to those obtained by Urry (1975) with the full parameterization process (2000 cases, 100 items). Urry's root mean square errors were .242, .123, and .056 for $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, respectively. At this level of accuracy, little shrinkage was in evidence. It should be borne in mind that, in the case of the on-line parameterization process, the number of cases can be increased at little or no cost. Also, as the quality of the bank is increases, more stringent termination rules can be introduced, further increasing accuracy of the on-line parameter estimates.

A final modification of the on-line parameterization process can be made which should further reduce estimation errors. As the parameterization procedure is presently set up, those examinees whose $\hat{\theta}$ do not attain the termination rule reliability within 30 items are dropped from the sample. Because coverage of θ is weakest in the Verbal Ability Bank in the low ranges, the dropped subjects tend to be concentrated in the low end of the distribution. This creates a paucity of information in a range in which many c_i values are determined, leading to higher c_i errors. Also, when the truncated distribution is restandardized, the result is a displacement of the $\hat{b}_i$ values. In the case of the Ideal Bank, no subjects were dropped at the .91 and .93 termination rules. Even at the .95 termination rule few examinees failed to reach the criterion (10 at N = 1000, 8 at N = 1500, and 9 at N = 2000). In the Verbal Ability Bank, no subjects were dropped at .91, but at .93, 23 were dropped at N = 1000, 53 at N = 1500, and 40 at N = 2000. Thus, up to 3.5% were eliminated. This probably explains to a great extent the failure of the .93 termination rule to produce noticeably better estimates than the .91 rule (Tables 4 and 5). Estimates would probably be improved by retaining in the sample those subjects who fail to reach the termination rule within 30 items. Although these $\hat{\theta}$ are less reliable, they probably provide information at low θ which is useful for parameterization purposes.

REFERENCES

Gugel, J., Schmidt, F. L., & Urry, V. W. *Effectiveness of the ancillary estimation procedure.* Paper presented at the Conference on Computerized Adaptive Testing, Washington, D.C., June 1975.

Jensema, C. J. *An application of latent trait mental test theory to the Washington pre-college test battery.* Unpublished doctoral dissertation, University of Washington, 1972.

Killcross, M. C. *A tailored testing system for selection and allocation in the British Army.* Paper presented at the 18th International Congress of Applied Psychology, Montreal, August 1974.

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

Urry, V. W. *Ancillary estimators for the item parameters of mental test models.* Personnel Research and Development Center, U.S. Civil Service Commission, 1975 in press (*a*).

Urry, V. W. *A five year quest: is computerized adaptive testing feasible?*Paper presented at the Conference on Computerized Adaptive Testing, Washington, D.C., June 1975 (*b*).

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