Item Selection Using an Average Growth Approximation of Target Information Functions

Richard M. Luecht and Thomas M. Hirsch
American College Testing

The derivations of several item selection algorithms for use in fitting test items to target information functions (IFs) are described. These algorithms circumvent iterative solutions by using the criteria of moving averages of the distance to a target IF and by simultaneously considering an entire range of ability points used to condition the IFs. The algorithms were tested by generating six forms of an ACT math test, each fit to an existing target test, including content-designated item subsets. The results indicate that the algorithms provided reliable fit to the target in terms of item parameters, test information functions, and expected score distributions. *Index terms: computerized testing, information functions, item information, parallel tests, test construction, test information.*

Advances in computer technology have generated a growing interest in test construction applications that take advantage of that technology. One such area of interest has been the use of computers to create parallel tests.

In item response theory (IRT), parallelism among tests, test forms, or subtests can be determined in part by item and test information functions (TIFs), among other criteria. IRT uses this concept of information, conditional on a latent ability, θ, to determine measurement precision. In contrast with classical test theory, which derives a single estimate of measurement accuracy through reliability and the standard error of measurement, IRT uses the inverse of the square root of the information function (IF) about θ to denote measurement accuracy across an entire θ metric.

This information is defined at the item level by

$$I_j(\theta) = \frac{P'(\theta)^2}{P(\theta)[1 - P(\theta)]} \quad , \tag{1}$$

where $P(\theta)$ is the probability of a correct response to item j at some θ level, and $P'(\theta)$ is the first derivative of $P(\theta)$. The item information, conditional on θ [$I(\theta)$], is additive over items. Consequently, information can be derived for an entire test or subtest as

$$T(\theta) = \sum_{j=1}^{J} I_j(\theta) \quad . \tag{2}$$

$T(\theta)$ is merely the TIF conditional on a single level of θ, but θ is distributed continuously from $-\infty$ to ∞. The shape of the TIF and its area can then be used to determine a weak form of parallelism among tests (Lord, 1977; Samejima, 1977). That is, tests or subtests having similar content, measuring the same latent trait, and having identical TIFs may be considered essentially parallel.

However, practical solutions that attempt to generate parallel tests through TIFs have demonstrated only limited success. Algorithms suggested by Theunissen (1985) and van der Linden and Boekkooi-Timminga (1989), which employ 0-1 linear programming to maximize test information, tend to require large amounts of computing time and are of limited utility in large-scale applications. Although parameter restrictions and heuristics can be applied to the 0-1 problem (e.g., Adema, 1988), there is a trade off between computer time and accuracy.

Techniques based on heuristic approaches (sort and search rule-based algorithms) dramatically reduce computational demands, but run the risk of operating with limited accuracy.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 1, March 1992, pp. 41-51

0146-6216/92/010041-11$1.80

For example, Ackerman (1989) was able to demonstrate the implementation of a strictly heuristic technique that prioritized item information based on distance from a target TIF. Under Ackerman's approach, pooled items were presorted at various θ levels by descending information, and those items that contributed the most information at priority points on the TIF were assigned to test forms. Unfortunately, Ackerman's technique tended to select the most discriminating items and usually overestimated the target TIFs (i.e., produced more informative tests than targeted).

Therefore, a set of techniques that effect a compromise between computational demands and heuristic approaches is necessary. This paper presents and evaluates a set of general heuristics and algorithms that can be used to select a prespecified number of items, J, from a pool of M items ($J < M$) that minimize the difference between a target IF and the actual IF formed by the J items, at K points on θ.

The Item Selection Algorithm

Derivation

$T(\theta_k)$ is defined as some amount of targeted test information, conditional on θ_k, for $k = 1, \ldots, K$ quadrature points along the θ metric. This target information is assumed to represent the standard form of a test. The properties of this standard form are to be matched. $T_j^*(\theta_k)$ is defined as the conditional estimated information with respect to the jth selected item ($j = 1, \ldots, J$) such that

$$\hat{T}(\theta_k) \equiv T_j^*(\theta_k) = \sum_{j=1}^{j^*} I_j(\theta_k) \quad . \tag{3}$$

By prior definition of the test information in Equation 2, $T_j^*(\theta_k)$ is an incremental sum of the conditional item information, $I_j(\theta_k)$. For conceptual convenience, $T_j^*(\theta_k)$, the approximation of the item information functions (IIFs) being incrementally summed, is distinguished from $\hat{T}(\theta_k)$, the finished approximation of the IF, conditional on θ_k. That is, $\hat{T}(\theta_k) = T_j^*(\theta_k)$, where $j = J$.

As implied earlier, the distribution of θ used to condition the TIF is generally considered to span $\{-\infty, \infty\}$; however, in practice, K is usually kept to some small number of quadrature points (e.g., $K \leq 31$) on the interval $\{-3.0, +3.0\}$ which is minimally adequate for sampling the cumulative information function (CIF, or cumulative density of the IF conditional on θ) at equal partitions. Because partitioning the information cumulative density function into equal areas essentially prioritizes the quadrature points of θ relative to the conditional information densities, the concentration and spread of θ corresponds closely to the actual distributional properties of the TIF.

Next, consider the distances between the target function, $T(\theta_k)$, and the IF under construction, $T_j^*(\theta_k)$. That distance is given by

$$d(\theta_k) = T(\theta_k) - T_j^*(\theta_k) \quad ,$$

where

$$d(\theta_k) = 0 \text{ for } T(\theta_k) \leq T_j^*(\theta_k) \quad , \text{ and}$$

$$d(\theta_k) = T(\theta_k) - T_j^*(\theta_k) \text{ for } T(\theta_k) > T_j^*(\theta_k) \quad , \tag{4}$$

which denotes the conditional difference between the target function, $T(\theta_k)$, and the approximation of the TIF, $T_j^*(\theta_k)$.

$d(\theta_k)$ can now be adjusted to a partitioned distance corresponding, ideally, to smooth growth in $T_j^*(\theta_k)$, given θ_k, as

$$\delta(\theta_k) = \frac{d(\theta_k)}{J - j + 1} \quad , \tag{5}$$

$$j = 1 \, , \ldots , J \quad .$$

This partitioning of the IF at some point, k, assumes that $\delta(\theta_k)$ is the optimal information with which to evaluate the next $J - j + 1$ items. Thus, $\delta(\theta_k)$ becomes a moving average of the information selection criterion and is adjusted at each iteration in the selection process.

There are two sound reasons for using $\delta(\theta_k)$.

First, the averaging process explicit in computing $\delta(\theta_k)$ appears to prevent extreme (and arbitrary) growth in any one area of the function. That is, items with maximal or minimal information properties at any kth θ point will be less likely to be selected than items with less extreme information. Thus, averaging should produce smooth growth in $T_j^*(\theta_k)$ as opposed to sporadic growth that requires continual and sometimes dramatic correction. Second, the dynamic nature of computing $\delta(\theta_k)$ at each jth selection iteration allows for constant "fine tuning" along the $\theta_k (k = 1, \ldots, K)$ points. In other words, error in estimating the target function is accounted for directly by the algorithm as part of the next set of distances from the target to be evaluated.

$\delta(\theta_k)$ is used to create a set of relative weights, $w(\theta_k)$, that will then be used to actually prioritize the information at the K θ points being evaluated. The relative weights are determined by normalizing $\delta(\theta_k)$ across the K quadrature points, as given by

$$w(\theta_k) = \frac{\delta(\theta_k)}{\sum_{k=1}^{K} \delta(\theta_k)} \quad , \tag{6}$$

where $\Sigma_{k=1}^{K} w(\theta_k) = 1.0$. In practice, $1 - w(\theta_k)$ will serve as the actual weight for reasons explained below.

$\delta(\theta_k)$ and $w(\theta_k)$ are used to evaluate the $M - j + 1$ items in the item pool. Let $\xi_m(\theta_k)$ denote the absolute error difference between the information of each mth item in the pool, evaluated at the kth θ point and $\delta(\theta_k)$. That is,

$$\xi_m(\theta_k) = | I_m(\theta_k) - \delta(\theta_k) | \quad , \tag{7}$$

where $\xi_m(\theta_k)$ is the error in fit of the $M - j + 1$ items in the unused item pool to $\delta(\theta_k)$, the partitioned IF. $\xi_m(\theta_k)$ is, in some sense, an arbitrary measure of the relative estimation error that occurs during the process of selecting items. Accordingly, rank ordering the absolute differences between $I_m(\theta_k)$ and $\delta(\theta_k)$, or squaring that difference, are suggested as plausible alternatives for arriving at $\xi_m(\theta_k)$. However, $\xi_m(\theta_k)$, only in its form as the absolute difference, retains the scale properties of the IFs under evaluation. Thus, any derivation of $\xi_m(\theta_k)$ except the use of the absolute difference would introduce additional, arbitrary, and probably unwanted weighting of the item information along the K θ points.

Finally, to determine the selection of the jth item, given $M - j + 1$ items, a composite selection value is created for each of the pooled items as a sum of each weighted relative error [i.e., a sum of the product of $1 - w(\theta_k)$ and $\xi_m(\theta_k)$, across the grid of K θ points]. Note that the use of $1 - w(\theta_k)$ in place of $w(\theta_k)$ merely guarantees that the weighting and the relative error in fit, $\xi_m(\theta_k)$, remain in the same direction. Summing the weighted relative errors produces an adjusted item selection composite of the fit to smooth growth in $T_j^*(\theta_k)$ for the $M - j + 1$ items remaining in the pool. That adjusted item fit selection composite, S_m, is given by the summation over the θ grid, where

$$S_m = \sum_{k=1}^{K} [1 - w(\theta_k)] \xi_m(\theta_k) \quad . \tag{8}$$

During each iteration of the selection cycle, the item with the smallest value of S_m (i.e., least overall error, weighted by information importance) is selected from the $M - j + 1$ pool, j is incremented, and the process continues until $j = J$ or until a specified degree of accuracy in approximating $T(\theta_k)$, $k = 1, \ldots, K$, is attained. Finding the item with the minimum value of S_m (per iteration) therefore serves as the primary heuristic to be used during the selection process.

Dealing with Item Subsets and Subtests

One assumption implicit in this algorithm is that the target function is comprised of fairly homogeneous items. That is, in building $T_j^*(\theta_k)$ (see Equation 3), the IFs are essentially compared to a criterion of an average IF for each of J items (conditional on the quadrature points, θ_k, $k = 1, \ldots, K$). In certain circumstances, this assumption may not be tenable. When a target

function is established as a composite of subsets of items from an existing test or from item specifications (e.g., subtests categorized by content area and/or some other criteria), the categorical subsets may have different information distributional properties (i.e., moments of the IFs) than the overall target IF.

In these situations, multiple targets can be used in a two-stage fitting procedure. The first stage of the method involves fitting each subtarget; the second stage groups the selected item subsets to fit an overall targeted TIF.

In the first stage of this procedure, a subtarget, $T_r(\theta_k)$, is fit, conditional on θ_k, comprised of J_r items for $r = 1, \ldots, R$ subsets of items, such that

$$T(\theta_k) = \sum_{r=1}^{R} T_r(\theta_k) \quad , \qquad (9)$$

$$k = 1 , \ldots , K \quad .$$

Thus, the subtarget represents an allowable partitioning of the IF in the overall target, given θ_k. In judging the fit of J_r items to $T_r(\theta_k)$, the item selection score, given by Equation 8, is now denoted as S_{rm} corresponding to the (restricted) subset of items in the pool. A subset of items, $T^*_{j_r}(\theta_k)$, is then independently fit to each $T_r(\theta_k)$ ($k = 1, \ldots, K$; $r = 1, \ldots, R$), where

$$T^*_{j_r}(\theta_k) = \sum_{j_r=1}^{J_r} I_{j_r}(\theta_k) \quad , \qquad (10)$$

$$k = 1 , \ldots , K \quad .$$

The second stage of fitting begins after all R subsets of J_r items have been fit to each subtarget. In this stage, the subsets of the J_r selected items are used as the basic units of comparison. The selection algorithm proceeds as described in Equation 8, but now compares the composite fit of the R subsets of selected J_r items, or $T^*_{j_r}(\theta_k)$, to the overall target $T(\theta_k)$. This item subset score is given by

$$S_{j_r} = \sum_{k=1}^{K} [1 - W(\theta_k)] \left| \sum_{j_r=1}^{J_r} I_{j_r}(\theta_k) - \delta(\theta_k) \right| \quad , \qquad (11)$$

where

$$\delta(\theta_k) = \frac{T(\theta_k) - \sum_{r=1}^{r^*} T^*_{j_r}(\theta_k)}{R - r + 1} \quad , \qquad (12)$$

with restrictions identical to those given in Equations 4 and 5, and where $W(\theta_k)$ is defined and used as shown in Equations 6 and 8. Therefore, the subset of J_r items that minimizes the weighted sum of information to the average growth in the overall conditional function being fitted is selected for each of the $r = 1, \ldots, R$ categories. It should be emphasized that R subsets of items (e.g., content areas) are guaranteed to be selected under this approach.

Multiple Parallel Test Forms

Multiple parallel test forms can be constructed in the same manner as a single test form. The major difference lies in the need to consider $T^*_{j_q}(\theta_k)$ ($j = 1, \ldots, J$; $q = 1, \ldots, Q$), where Q is the number of test forms being fit to the target, $T(\theta_k)$. By rotating the order of the form being fit (q) at each jth item selection iteration and by controlling for duplication of item selection across forms, the assignment of items—based on the information fit to $\delta_q(\theta_k)$—can be essentially equalized across test forms.

Method

Implementation

All algorithms and heuristics were implemented in an IBM-compatible microcomputer-based package called ITEMSEL. This integrated software consists of 10 menu-driven program modules written in Microsoft QuickBasic 4.0 (1987) by the first author. ITEMSEL provides graphic on-screen presentation of the selection process and a wide variety of item database modules and file handling utilities that facilitate the item selection process. The software package also supports the construction of multiple test forms, the use of multiple subtargets for dealing with content subtests or subsets of items, and allows user submitted item substitutions.

ITEMSEL requires user input of an item pool file, a target information file, related control

input such as the size of J or $J(r)$ (the number of items to be selected), and classifications of the items to be selected. Selected items are retained in additional files in which optimization of the fitting process can occur or from which optional combining of item subtests can be accomplished.

ITEMSEL uses a three-parameter logistic IRT model to compute all information quantities. Under that model, the probability of a correct response to item j, conditional on θ, is given by

$$P_j(\theta) = c_j + (1-c_j)\ \{1 + \exp[-Da_j(\theta - b_j)]\}^{-1}\ , \quad (13)$$

where c_j is the lower asymptote parameter, a_j is the discrimination parameter, and b_j is the item difficulty. D is a constant equal to approximately 1.702 and is used for scaling θ under the logistic model.

Data Specifications

An item pool consisting of 600 mathematics items from American College Testing (ACT) mathematics tests was selected to investigate the use of the S_m and $S_{m'}$ algorithms as implemented by the ITEMSEL program. 520 of the items were from 13 previously administered ACT Assessment Program (AAP) Math Usage tests. An additional 80 items were drawn from the Collegiate Mathematics Placement Program. Item parameters for all 600 items were derived from a three-parameter logistic calibration performed using LOGIST IV (Wingersky, Barton, & Lord, 1982) and were scaled to a common θ metric.

40 items that comprised the AAP Math Usage Form 26A were selected as the overall test target function, to remain consistent with Ackerman (1989). These 40 items were also included in the item pool. The Form 26A target function was fit by evaluating the test information at $K = 31$ quadrature points on the θ interval $\{-3.0, +3.0\}$. The decision to use 31 quadrature points was based on previous analyses in which stable and satisfactory results were obtained with minimal computing time and computer memory requirements. A smaller number of quadrature points might, of course, yield nominally different item selections. The CIF was equally partitioned (based on an integration of 1,000 θ points) to locate the 31 points, which were selected so that they divided the CIF into equal area partitions.

Additionally, the six content areas that comprise Form 26A of the AAP Math Usage test were used to generate six corresponding subtargets. The CIF of each subtarget was likewise partitioned independently when generating the $K = 31$ quadrature points. For purposes of computing the IFs and generating subsequent subsets of items, these Form 26A subtest content areas contained the following numbers of items: Arithmetic and Algebraic Reasoning (AAR), 14 items; Arithmetic and Algebraic Operations (AAO), 4 items; Geometry (G), 8 items; Intermediate Algebra (IA), 8 items; Number and Numeration Concepts (NNS), 4 items; and Advanced Topics (AT), 2 items.

Item Selection Procedures

ITEMSEL was employed in a two-stage set of fitting procedures meant to generate six independent forms of the AAP Math Usage test. In the first stage, six forms of each of the content areas (AAR, AAO, G, IA, NNS, and AT) were initially fit to the Form 26A subtest information targets. ITEMSEL thus generated a total of 36 content-restricted item subsets. In the second stage of fitting, an "optimizer" module in the ITEMSEL system was used to identify and combine composite groupings of the content-restricted item subsets that fit the overall Form 26A target IF to produce six independent forms of the AAP Math Usage test. That is, each of the six generated total test forms was created as a summation of the unique AAR, AAO, G, IA, NNS, and AT subsets of items that "best" fit the overall Form 26A target function.

The generation of multiple forms during both stages of item selection was performed as a simultaneous operation. As described earlier, ITEMSEL automatically rotated all form indices as each item or item subtest was selected to ensure equalization of the item/subtest selection process across forms.

Table 1
Descriptive Statistics for Fitted IRT Item Parameters for Form 26A Target Test and AAP Math Forms A-F (N = 40 items)

Test Form	Mean			SD			Skewness			Kurtosis		
	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
26A	1.03	.29	.16	.40	.60	.04	.92	-.62	.03	.19	-.47	1.19
A	1.03	.35	.17	.29	.52	.06	.87	-.20	.52	.96	-.88	.20
B	1.05	.35	.17	.29	.50	.06	1.03	-.31	2.17	1.68	-.79	9.65
C	1.05	.31	.17	.30	.54	.05	.71	-.12	.25	1.09	-1.06	-.25
D	1.05	.32	.16	.29	.55	.06	1.46	-.11	.81	2.56	-.66	1.50
E	1.04	.31	.17	.28	.50	.06	1.25	-.49	-.32	1.88	-.64	.09
F	1.01	.32	.15	.29	.50	.05	.68	.13	-.27	.88	-.94	.27

Results

IRT Item Parameters

The IRT item parameters provide an important starting point in consideration of the item selection process. Assuming that the test target represents an ideal composite of items, the items selected or fitted by the ITEMSEL program should demonstrate distributions of the item parameters similar to those present in the target specifications or test.

Table 1 compares the distributional properties of the parameters for each of the six generated AAP Math forms (A-F) with the Math Form 26A target parameters. The data suggest a very slight tendency (with one exception, Form F) toward overfitting the average item *a* parameters, and toward selecting items with nominally higher mean *b* parameters. At the same time, the SDs of both the discrimination parameters and the item difficulty parameters for the fitted forms (A-F) were slightly smaller than for the Form 26A target test.

Thus, there appears to be a tendency for ITEMSEL to spread out the information (i.e., to slightly underfit at the peak of the IF and compensate elsewhere along the function, at least for these data). Given the explicit averaging of the conditional IFs, through the S_m algorithm, this very minor distributional difference seems quite reasonable. It should also be noted that despite the minor distributional differences between the item parameters of the target test and those of the selected test forms, ITEMSEL was nonetheless very consistent in matching item parameters among Forms A through F of the test.

Table 2
Means and SDs of IRT Item Parameters for 12 Manually Constructed AAP Math Usage Forms (N = 40 Items)

Test Form	*a*		*b*		*c*	
	Mean	SD	Mean	SD	Mean	SD
24B	1.058	.296	.309	.661	.160	.084
25B	.994	.247	.395	.973	.159	.079
25C	1.078	.379	.359	.744	.157	.077
25D	1.068	.353	.321	.830	.142	.079
25E	1.057	.259	.307	.633	.128	.055
25F	.950	.370	.385	.863	.152	.062
26B	.989	.358	.240	.875	.172	.046
26C	.930	.365	.328	.876	.162	.034
26D	.951	.427	.392	1.283	.185	.026
26E	.972	.297	.254	.777	.166	.048
26F	.926	.365	.342	.953	.159	.034
27A	.990	.394	.332	.868	.178	.046

Table 2 shows the means and SDs of the IRT parameters from 12 manually constructed Math Usage forms (i.e., actual forms prepared by test specialists). These data provide strong evidence of a greater degree of variation in the types of items that were manually selected across forms than was present in the computer-selected forms.

Information Functions

Figure 1 shows that all six selected Math Usage forms demonstrated quite similar patterns of information. That similarity is perhaps even more evident in terms of the means and variances of the IFs (for which estimates of the expectations can be derived across the 31 quadrature points of θ). For the target test, Form 26A, the mean information across the 31 quadrature points of θ was 21.67. In comparison, the average of the expected means of the TIFs for the six selected test forms was 21.54. Likewise, the approximate variance of the Form 26A target IF for 31 quadrature points was 152.53, compared to an average variance of 161.55 for the Forms A–F TIFs. These results indicate that the IFs from the six selected test forms were essentially centered at the same point as the target function, but with nominally larger variances.

Figure 1
Test Information Curves for Six Forms of the AAP Math Usage Test Fit to the Form 26A Target Information Curve

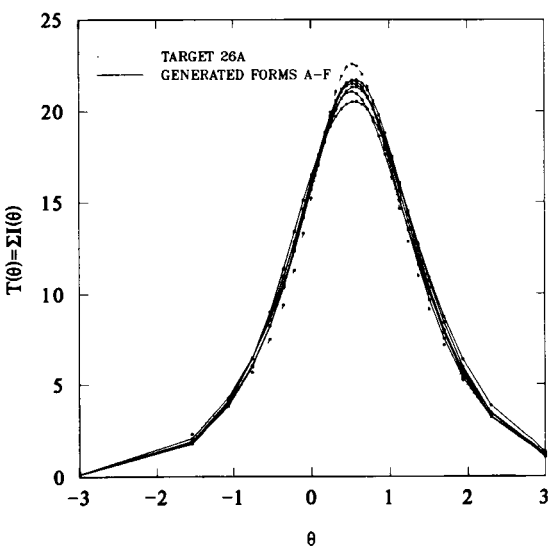


Figure 2 shows IFs for the subsets of items selected by ITEMSEL to fit the individual content area subtargets (AAR, AAO, G, IA, NNS, and AT). Some caution is warranted, however, when reviewing these content-specific graphs of the item subsets. The apparent differences in the functions across content areas must take into account the scaling of the ordinate axes. For example, the AT forms (Figure 2f) appear to demonstrate a greater lack of fit than the AAR forms (Figure 2a). However, by considering the ordinate axes of the AT functions versus the AAR functions, it should be obvious that the real differences between the AT functions (two items per subtest form) are actually as small or smaller than the differences between the AAR functions (14 items per subtest form).

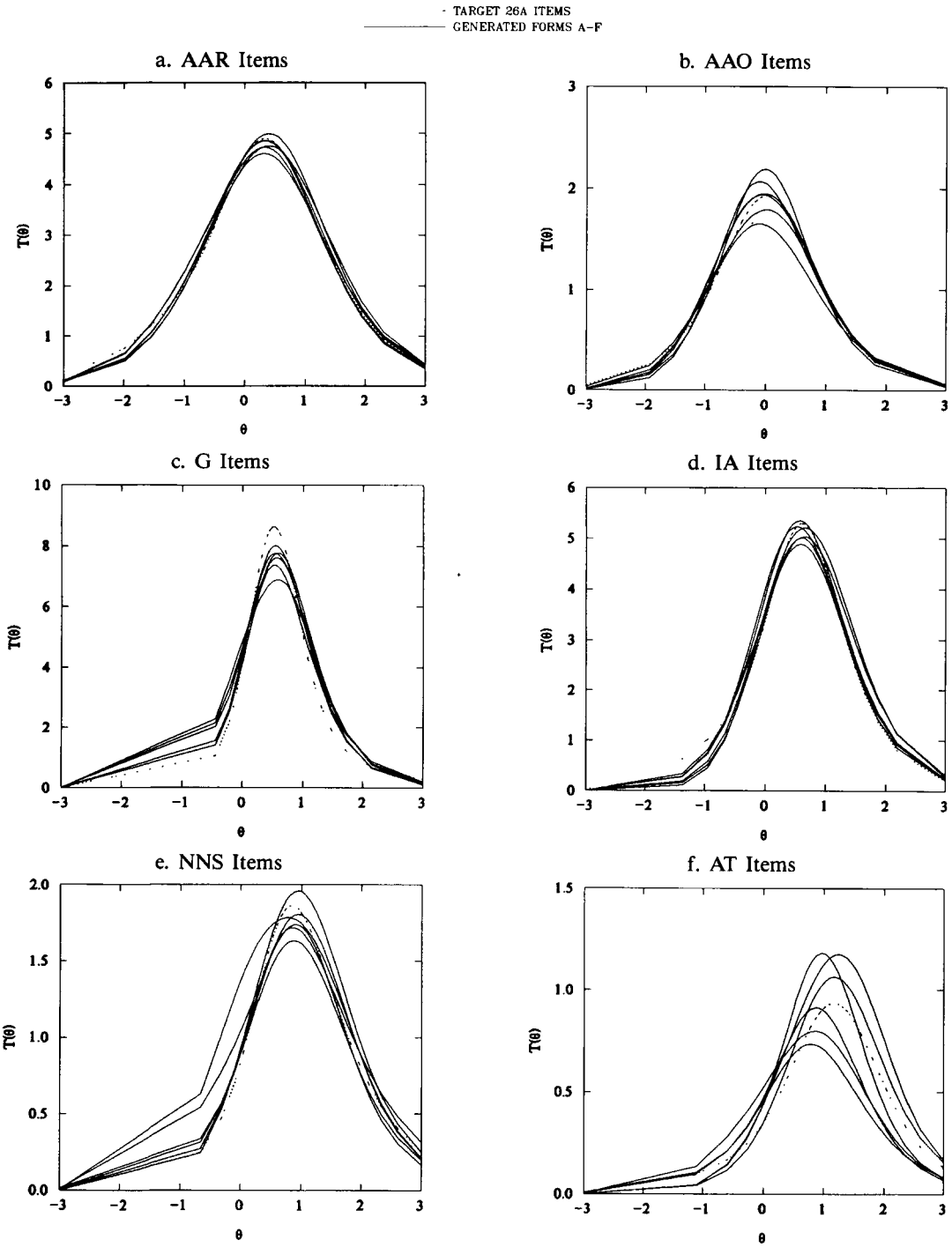
Goodness of Fit

The weighted mean square (WMS) shown in Table 3 represents a weighted difference index, where the weights are derived from the test characteristic function associated with the Form 26A target test. That is, the weights are used to scale the squared IF differences to an assumed distribution of true scores for the target test. (Here, the term "true score" denotes the test characteristic function conditional on some range of θ.) For the present AAP data, those true scores were assumed to be normally distributed. Under this normality assumption, the actual weighting

Table 3
WMS, PMI, and $WMS_{SE(\theta)}$ Goodness-of-Fit Indices for Fit of Generated Forms to Form 26A

Test Form	WMS	PMI	$WMS_{SE(\theta)}$
A	.874	.040	.006
B	.637	.029	.011
C	1.018	.047	.004
D	.733	.034	.005
E	.655	.030	.005
F	1.885	.087	.002

Figure 2
Test Information Curves for the Six Form 26A Content Classifications



was accomplished by standardizing the true scores across a range of 31 equidistant quadrature points for θ between ±3.0, and then using a normal density approximation function to compute the weights. The net effect of this procedure was to assign the greatest weight to squared information differences near the point of maximum slope of the target test response function (TRF) and the smallest weight near the asymptotes.

This WMS statistic is given by

$$\text{WMS} = \sum_{k=1}^{K} \Phi[\zeta(\theta_k)]\,[T(\theta_k) - \hat{T}(\theta_k)]^2 \quad , \qquad (14)$$

where

$$\Phi[\zeta(\theta_k)] = \frac{\exp[-.5\zeta(\theta_k)^2]}{\sum_{k=1}^{K} \exp[-.5\zeta(\theta_k)^2]} \qquad (15)$$

and

$$\zeta(\theta_k) = \frac{\sum_{j=1}^{J} P_j(\theta_k) - \left[\sum_{k=1}^{K}\sum_{j=1}^{J} P_j(\theta_k)\right] / K}{\left[\frac{K\sum_{k=1}^{K}\sum_{j=1}^{J} P_j(\theta_k) - \left[\sum_{k=1}^{K}\sum_{j=1}^{J} P_j(\theta_k)\right]^2}{K(K-1)}\right]^{1/2}} \qquad (16)$$

given $P_j(\theta_k)$, the probability of a correct response to item j, conditional on θ_k.

Typically, comparisons of item response functions and IFs are made using a simple, unweighted mean square statistic. However, a mean square statistic does assume uniform weights. Unfortunately, it is also highly susceptible to large differences near the asymptotes of the functions as well as to the range of θ used during computation. The use of the true score weighting, under the assumption of normality, is essentially a compromise between implicitly using such uniform weights, explicitly assuming some arbitrary distribution of the unobservable θ (e.g., normal 0,1), and directly using the normalized target test information to weight the difference.

By itself, the WMS goodness-of-fit index provided in Table 3 implies a weighted function of the squared differences between the Form 26A target function and the selected TIFs (i.e., the functions for Forms A–F). However, to put this index in a different perspective, consider it a proportion of an IF, conditional on some value of θ. To do so merely requires dividing the value of the WMS index in Table 3 by the IF at some point along the θ metric (e.g., the mean information for the Form 26A target test of 21.67). For example, the WMS value .874 in the first row of Table 3 could be seen to represent proportional differences between the Form A function and the target function of 4.03%, at the point of average test information. These proportional differences, conditional on the mean information in the Form 26A target function, are also provided in Table 3 as PMI, the proportion of mean information. The PMI results indicate that the fit between the IFs is actually better than the WMS indices might suggest on the surface. That is, the apparent functional differences taken as relative ratios (proportions) to the amount of average information in the target function are essentially inconsequential.

Another method of assessing the goodness of fit considers the relationship between the test information and the standard error of the latent abilities, θ:

$$\sigma_e(\theta) = \left[\sum_{j=1}^{J} I_j(\theta)\right]^{-1/2} \quad . \qquad (17)$$

Using this relationship, it becomes possible to restate the goodness-of-fit statistics as weighted functions of the average unsigned differences between the conditional standard errors. These standard error differences are also shown in Table 3. Considering these weighted differences between the standard errors, the observed differences between the conditional target test and the generated TIFs obviously become quite negligible.

Expected Score Differences

The final determinants of the adequacy and accuracy in fitting a target test using S_m and S_{rm} algorithms (as implemented in the ITEMSEL software) are the expected score distributions obtained from the various tests. That is, if the issue

of parallelism among test forms is considered to extend beyond the objective function (test information), then the score distributions of the fitted test forms in comparison to the target test (AAP Math Form 26A, in this case) must also be considered.

Figure 3
TRFs for the Target Test Form 26A and Six Test Forms Fitted by ITEMSEL

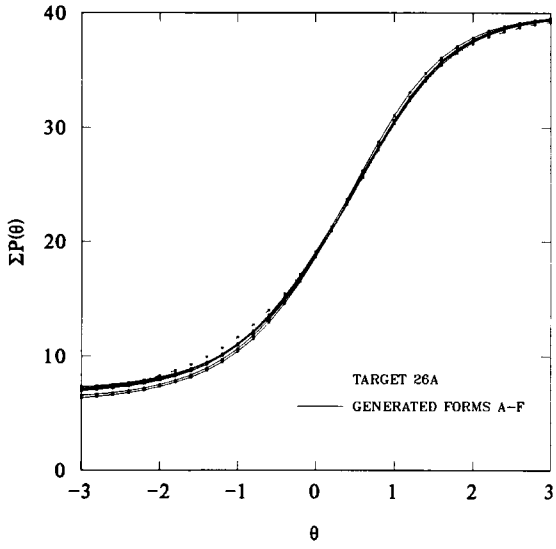


Figure 3 presents the TRFs for each of the six fitted test forms and the TRF for Form 26A. These TRFs are defined by the sum of the conditional probabilities for all items in a test across the θ metric. That is,

$$\psi(\theta) = \sum_{j=1}^{J} P_j(\theta) \quad . \tag{18}$$

Therefore, $\psi(\theta)$ defines the expectation of a random examinee's true score on J items, given his/her θ level (Lord, 1980).

Figure 3 demonstrates a very close correspondence between true scores across the fitted forms of the AAP Math test and Form 26A. Additionally, the differences between predicted score distributions can be compared by converting the true scores to a discrete number-correct (NC) scale, using a compound binomial generating function, where the density of θ is assumed (Lord, 1980). Predicted NC scores were obtained by assuming a (0,1) normal distribution on θ. Table 4 provides the means, SDs, skewness, and kurtosis values of the predicted score distributions for the six AAP Math test forms fitted by ITEMSEL and the Form 26A target test. Classical item difficulties (proportion correct), biserial correlations, and their SDs are also shown in Table 4.

Table 4 provides evidence of essential parallelism among the six fitted forms and the target test, not only in terms of predicted means and SDs, but also skewness and kurtosis. In other words, the process of fitting the target information was sufficient to fit the expected (predicted) NC score distributions for the present item pool. Finally, the data in Table 4 suggest that the S_m and S_{rm} algorithms also satisfy classical test theory criteria for parallelism.

Discussion

The S_m and S_{rm} algorithms have three distinct benefits. First, the moving average criterion absorbs and redirects error in fit, thus allowing

Table 4
Descriptive Statistics for Proportion-Correct (PC), Biserial r (r_{bis}), and NC Scores for Six Fitted Test Forms and Target Form 26A

Test	PC		(r_{bis})		NC Scores			
Form	Mean	SD	Mean	SD	Mean	SD	Skew	Kurtosis
26A	.495	.126	.591	.079	19.825	8.937	.369	−.812
A	.496	.117	.585	.065	19.840	8.926	.361	−.808
B	.493	.117	.586	.070	19.734	8.959	.365	−.844
C	.492	.128	.588	.064	19.686	8.913	.331	−.826
D	.497	.128	.598	.070	19.874	9.071	.333	−.845
E	.489	.102	.594	.070	19.551	9.171	.337	−.878
F	.503	.111	.594	.081	20.139	9.117	.330	−.846

for a noniterative solution. The result is a reasonably fast method of fitting any target IF. Second, the algorithms simultaneously consider all quadrature points that define the TIFs and on which the IFs are conditional. That is, the entire IF is always fit in the process of selecting items or item subsets. Finally, the algorithms can be conveniently extended for use with subtests/subtargets, item subsets, and multiple test forms.

ITEMSEL was able to produce six test forms that were essentially parallel (although not strictly parallel) to the target test, as evaluated by multiple criteria. For example, IRT item parameters were shown to closely correspond to the parameters in the target test—more closely, in fact, than the parameters derived from existing, manually constructed forms of the Math Usage test. Other criteria denoting the fit of the selected test forms to the target test (e.g., comparisons of the actual IFs) likewise demonstrated consistency among the generated forms. In addition, the procedure generated similar expected score distributions and classical item parameters. Finally, the method is feasible for microcomputer technology, and is at least as accurate as manual test construction methods.

References

Ackerman, T. A. (1989, April). *An alternative methodology for creating parallel test forms using the IRT information function.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Adema, J. J. (1988). *A note on solving large-scale zero-one programming problems* (Research Rep. No. 88-4). Twente: University of Twente, Netherlands, Department of Education.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117–138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Microsoft Corp. (1987). *Quick Basic 4.0* [Computer program]. Redmond WA: Microsoft Press.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika, 42,* 193–198.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50,* 411-420.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54,* 237–247.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton NJ: Educational Testing Service.

Author's Address

Send requests for reprints or further information to Richard M. Luecht, Research Psychologist, Support, Tech. Applications & Research, American College Testing, P.O. Box 168, Iowa City IA 52243, U.S.A.