ACT Research Report Series 90-6

Computerized Test Construction Using an Average Growth Approximation of Target Information Functions

Richard M. Luecht
Thomas M. Hirsch

August 1990

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

COMPUTERIZED TEST CONSTRUCTION USING AN AVERAGE GROWTH APPROXIMATION OF TARGET INFORMATION FUNCTIONS

Richard M. Luecht
Thomas M. Hirsch

American College Testing

June, 1990

ABSTRACT

This paper describes the derivation of several item selection algorithms for use in fitting test items to target information functions. These algorithms circumvent iterative solutions by using the criteria of moving averages of the distance to a target information function and simultaneously considering an entire range of ability points used to condition the information functions. The algorithms were implemented in a microcomputer software package and 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 provide reliable fit to the target in terms of item parameters, test information functions and expected score distributions. A discussion of the application is included.

Introduction

Advances in computer technology have generated a growing interest in test construction applications which 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 in part be determined by what are termed *item* and *test information functions* among other criteria. IRT uses this concept of information, conditional upon a latent ability, θ, to determine measurement precision. Contrasted with classical test theory, which derives a single estimate of measurement accuracy via reliability and the standard error of measurement, IRT uses the inverse of the square root of the information function about the θs to denote measurement accuracy across an entire latent ability metric.

This *information* is defined at the item level by

$$I_j(\theta_k) = \frac{P_j'(\theta_k)^2}{P_j(\theta_k)Q_j(\theta_k)} \tag{1}$$

where $P_j(\theta_k)$ is the probability of a correct response to item j at some ability level, θ_k, $Q_j(\theta_k) = 1 - P_j(\theta_k)$, and $P_j^{'}(\theta_k)$ is the first derivative of $P_j(\theta_k)$ with respect to θ_k. Furthermore, the item information, $I_j(\theta_k)$, is additive which allows us to derive the information for an entire test or subtest as

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

It must be noted, however that $T(\theta_k)$ is merely the test information function conditional upon some single level of ability, θ_k. Because the θ_k abilities are in reality distributed continuously on R or the real number line, {-∞, +∞}, we must extend our concern beyond some k^{th} ability point to an entire *test information curve*. The shape of and area under such test information curves can then be used to determine a weak form of parallelism among tests (Lord, 1977, Samejima, 1977). That is, tests (forms or subtests having similar content and measuring the same latent trait) with identical test information curves may be considered essentially to be *parallel*. Therefore, if we can create different test forms with approximately the same test information curves (and similar content), then our forms should be reasonably parallel.

However, practical solutions to the problem of actually generating parallel tests via test information curves have demonstrated only limited success. Algorithms suggested by Theunissen (1985) and van der Linden and Boekkooi-Timminga (1989), which employ zero-one, linear programming to maximize test information, tend to require large amounts of computing time and remain limited for large scale applications. Although parameter restrictions and heuristics can be applied to the zero-one problem (e.g. Adema, 1988) a trade-off of computer time versus accuracy tends to result.

Other techniques based upon more heuristic approaches (sort and search rule-based algorithms) more dramatically reduce computational loads but run the risk of operating with limited accuracy. For example, Ackerman (1989) was able to demonstrate the implementation of a strictly heuristic technique which prioritized item information based upon distance from a target test information curve. Under Ackerman's approach, pooled items were presorted at various ability levels by descending information and those items which contributed the most information at priority points on the test information curve were assigned to test forms. Unfortunately, Ackerman's technique tended to always choose the most discriminating items and usually overestimated the target test information curves (i.e. produced more informative tests than targeted).

What appears necessary, therefore, is a set of techniques which effect a compromise between computational loads and purely heuristic approaches. This paper focuses upon that specific problem--to determine a set of general heuristics and algorithms which can be used to select J items from a pool of M items (J<M) which minimize the difference between a target information curve and the actual information curve formed by the J items, at some K points along an ability distribution.

Derivation of the Item Selection Algorithm

We begin by defining T_k as some amount of *targeted* test information, conditional upon θ_k, (k = 1,..., K quadrature points). This target information is assumed to represent the standard form of a test whose properties we wish to match. We also define T^*_{jk} as the conditional information with respect to the jth selected item (*j = 1,..., J, k = 1,..., K) such that

$$\hat{T}_k = T^*_{jk} = \sum_{j=1}^{j*} I_j(\theta_k) \; . \qquad (3)$$

Note that by prior definition of the test information, equation (2), T^*_{jk} is merely an incremental sum of the item information, $I_j(\theta_k)$. To further clarify equation (3), it is only for conceptual convenience that we distinguish between T^*_{jk} as the approximation of the item information functions being incrementally summed and $\hat{T}_k$ as the finished approximation of the information function, conditional upon θ_k (i.e. $\hat{T}_k = T^*_{jk}$ where j = J).

As implied earlier, the ability distribution of θs used to condition the test information curve is generally considered to span {-∞, +∞}; however, in practice, K is usually kept to some small number of quadrature points (e.g. K < = 31) on the interval {-3.0, +3.0}) minimally adequate for sampling the cumulative information function (CIF, or cumulative density of the information function conditional on θ) at equal partitions.

Next, we need to consider the distances between the target function, T_k, and the information function under construction, T^*_{jk}. That distance is given by

$$d_k = |T_k - T^*_{jk}| \left\{ \begin{array}{l} d_k = 0 \text{ for } T_k \leq T^*_{jk} \\ d_k = T_k - T^*_{jk} \text{ for } T_k > T^*_{jk}, \end{array} \right. \qquad (4)$$

which denotes the absolute difference (distance) between the target function, T_k, and the approximation of the test information function, T^*_{jk}.

We can now adjust d_k to a partitioned distance corresponding, ideally, to smooth growth in T^*_{jk}, given θ_k, as

$$\delta_k = \frac{d_k}{J - j + 1}, \; j = 1..., J. \qquad (5)$$

This partitioning of the information function at some point, k, assumes that δ_k is the optimal information with which to evaluate the next J - j + 1 items. In short, δ_k becomes a moving average of the *information selection criteria* and is adjusted at each iteration in the selection process.

There appear to be two sound reasons for using δ_k. First, the *averaging* process explicit in computing δ_k would appear to prevent extreme (and arbitrary) growth in any one area of the curve. That is, items with maximal or minimal information properties at any k^{th} ability point will be less likely to be chosen than items with less extreme information. Thus, averaging should produce smooth growth in T^*_{jk} as opposed to sporadic growth which requires continual and sometimes dramatic correction. Second, the *dynamic* nature of computing δ_k at each j^{th} selection iteration allows for constant "fine tuning" along the θ_k (k = 1...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.

Once δ_k is derived, we use it to create a set of relative weights, ω_k, which will then be used to actually prioritize the information at K ability points being evaluated. The relative weights are determined by normalizing the δ_k's across the k quadrature points, as given by

$$\omega_k = \frac{\delta_k}{\sum_{k=1}^{K} \delta_k} \quad , \tag{6}$$

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

We now proceed to use δ_k and ω_k to evaluate the M - j + 1 items in the item pool. Let ξ_{mk} denote the absolute error difference between the information of each m^{th} item in the pool, evaluated at the k^{th} ability point, and δ_k. That is,

$$\xi_{mk} = |I_{mk} - \delta_k| \quad , \tag{7}$$

where ξ_{mk} might be called the *error in fit* of the M - j + 1 items in the unused item pool to δ_k. It should be noted that in some sense δ_{mk} is an arbitrary measure of the relative estimation error during the process of selecting items. Accordingly, rank ordering the absolute differences between I_{mk} and δ_k or squaring that difference might each be suggested as plausible alternatives for arriving at ξ_{mk}. However, *only* ξ_{mk} in its form as the absolute difference retains the scale properties of the information functions under evaluation. In short, any derivation of ξ_{mk} *except* by using the absolute difference would introduce additional, arbitrary and probably unwanted weighting of the item information along the K ability points.

Finally, to determine the selection of the j^{th} item, given $M - j + 1$ items, we need to create a composite selection value for each of the pooled items as a sum of each weighted relative error (i.e. a sum of the product of $1 - \omega_k$ and ξ_{mk}), across the K ability points. Note that the use of $1 - \omega_k$ in place of ω_k merely guarantees that the weighting and the relative error in fit, ξ_{mk}, remain in the same direction. By summing the weighted relative errors, we arrive at an adjusted item selection composite (of the fit to smooth growth in T^*_{jk}) for the $M - j + 1$ items remaining in the pool. That adjusted item fit selection composite, S_m, is given by

$$S_m = \sum_{k=1}^{K} (1 - \omega_k)\xi_{mk} \quad . \tag{8}$$

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 chosen 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_k ($k = 1...K$) 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 the algorithm described in the prior section is that the target curve is comprised of fairly homogeneous items. That is, in building T^{*}_{jk} (see equation [3]), the item information functions are essentially compared to a criterion of an average information function for each of J items (conditional upon the quadrature points, θ_k, k = 1...K). In certain circumstances, this assumption may not be tenable. Where a target curve 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 information curves, than the overall target information function.

In these situations, multiple targets can be used in a two-stage fitting procedure. Essentially, the method involves fitting each categorical or criterial subtarget in the first stage and then grouping the selected item subsets in a second stage to fit an overall targeted test information function.

In the first stage of this procedure, we presume to fit a subtarget, T_k, conditional upon θ_k, comprised of $J_{(r)}$ items for r=1...R subsets of items such that

$$T_k = \sum_{r=1}^{R} T_{rk},\ k = 1...K \qquad (9)$$

Thus, the subtarget represents an allowable partitioning of the information function in the overall target, given θ_k. In judging the fit of $J_{(r)}$ items to the subtarget, T_{kr}, the item selection score, given by equation (8), is now denoted as S_{rm} corresponding to the [restricted] subset of items in the pool. We then independently fit a subset of items, $T^*_{J(r)k}$, to each T_{rk} subtarget (k = 1...K, r = 1...R), where

$$T^*_{J(r)k} = \sum_{j_{(r)}=1}^{J_{(r)}} I_{j(r)}(\theta_k), \; k = 1...K \qquad (10)$$

After all R subsets of $J_{(r)}$ items have been fitted to each subtarget, T_{rk}, we proceed to the second stage of fitting. In this stage, we use the subsets of the $J_{(r)}$ selected items 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 $T^*_{J(r)k}$, to the overall target T_k. This item subset score is given by

$$S_{J(r)} = \Sigma \, (1 - W_k) \left| \sum_{j_{(r)}=1}^{J_{(r)}} I_{j(r)}(\theta_k) - \delta_k \right| \qquad (11)$$

where

$$\delta_k = \frac{T_k - \sum\limits_{r=1}^{r} T^*_{J(r)k}}{R - r + 1} \qquad (12)$$

with restrictions identical to those given in equations (4) and (5), and where W_k is defined and used as shown in equations (6) and (8). Therefore, the subset of $J_{(r)}$ items which minimizes the weighted sum of information to the average growth in the conditional curve being fitted is selected for r = 1...R cycles.

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^*_{jkq} (j = 1...J, k = 1...K, q = 1...Q), where Q is the number of test forms being fit to the target, T_k. Furthermore, by rotating the order of the form being fit (q) at each j^{th} item selection iteration and controlling for duplication of item selection across forms, the assignment of items (based upon their information fit to δ_{kq}) can be essentially equalized across test forms.

Methods

Implementation

All algorithms and heuristics discussed in the prior section were formally 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 features EGA/VGA graphics for on-screen presentation of the selection process and provides a wide variety of item data base modules and file handling utilities which facilitate the item selection process. The software package also fully supports the construction of multiple test forms, the use of multiple subtargets for dealing with content subtests or subsets of items and even allows user submitted item substitutions.

The basic process of using ITEMSEL involves user inputs of an item pool file, a target information file, related control inputs such as the size of J or $J_{(r)}$ (the number of items to be selected) and content filter values/text. Selected items are retained in additional files where optimization of the fitting process can occur or from which optional combining of item subtests can be accomplished.

The programs assume a 3-parameter IRT or logistic model for purposes of computing all information quantities. Under that model, the probability of a correct response to item j, conditional on ability, θ_k, is given by

$$P_j(\theta_k) = c_j + (1 - c_j)\{1 + e^{-Da_j(\theta_k - b_j)}\}^{-1} \qquad (13)$$

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 used for scaling θ under the logistic model.

Data Specifications

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

40 items which comprised the AAP Mathematics Form 26A were selected as the overall test target curve to remain consistent with a previously noted study conducted by Ackerman (1989). These 40 items were also included in the item pool. The Form 26A target curve was fit by evaluating the test information at K = 31 quadrature points on the θ interval $\{-3.0, + 3.0\}$. The cumulative information function (CIF) was equally partitioned (based upon an integration of 1000 θ points) to locate the 31 points. That is, points were selected which divided the information curve into equal area partitions.

Additionally, the six content areas which comprise Form 26A of the AAP Mathematics test were used to generate six corresponding subtargets. The CIF of each subtarget was likewise partitioned independently when generating the K = 31 quadrature points. These Form 26A subtest content areas contained the following numbers of items (for purposes of computing the information functions and generating subsequent subsets of items): AAR = 14 items, AAO = 4 items, G = 8 items, IA = 8 items, NNS = 4 items and AT = 2 items.

Item Selection Procedures

The ITEMSEL microcomputer program was employed in a two-stage set of fitting procedures meant to generate six independent forms of the AAP Mathematics 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 which fit the overall Form 26A target information curve to produce six independent forms of the AAP Mathematics test (see Dealing with Item Subsets and Subtests). That is, each of the six generated total test forms was created as a summation of the unique AAR, AAO, G, IA, NSS and AT subsets of items which "best" fit the overall Form 26A target curve.

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.

Results

In the present study, six forms of 40 items each were generated by ITEMSEL using the 600 items in the math pool and the Mathematics test Form 26A target information values conditional on K=31 quadrature points of θ. In assessing the quality of the algorithms to fit the Form 26A target a number of considerations and comparisons are presented.

Summary of IRT Item Parameters

The IRT item parameters (discrimination, difficulty and the lower asymptote) provide an important starting point in consideration of the item selection process. Assuming that the test target represents an ideal composite of items, we would expect that the items selected or fitted via the ITEMSEL program should demonstrate similar distributions of the item parameters to those present in the target specifications or test.

A summary of the means and standard deviations of the item parameters is presented in Table 1. This table compares the distributional properties of the parameters for each of the six generated AAP Mathematics test Forms (A-F) with the Mathematics test Form 26A target parameters. In general, the apparent trend of the parameters suggests a very slight tendency (with one exception, Form F) by ITEMSEL toward overfitting the average item discrimination parameters (a) and toward choosing items with nominally higher mean difficulty parameters (b).

Insert Table 1 about here

The net result appears to be, therefore, a tendency for ITEMSEL to spread out the information (i.e. produce a more platykurtic distribution of information). Given the explicit averaging of the conditional information functions, via the S_m algorithm, this 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.

As an additional comparison, consider Table 2 which shows the means and standard deviations of the IRT parameters from 12 manually constructed Mathematics test Forms (i.e. actual forms prepared by ACT test development staff). Table 2 would appear to provide strong evidence of a greater degree of variation in the types of items which were manually selected across forms than was present in the computer-selected forms summarized in Table 1. It should be noted, however, that these 12 manually-constructed test forms did not use target test information as the objective criteria.

Insert Table 2 about here

Goodness-of-Fit

In addition to the descriptive summary of the item parameters, we can also consider the test information curves, themselves. As shown in Figure 1, all six selected Mathematics Test Forms (A-F, 40 items each) demonstrated quite similar patterns of information. That similarity is perhaps even more evident in terms of the means and variances of the information curves (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. Comparatively, the average of the expected means of the test information curves for the six selected test forms (A-F) was 21.54. Likewise, the approximate variance of the Form 26A target information curve for 31 quadrature points was 152.53. This compares to an average variance of 161.55 for the Form A-F test information curves. Therefore, the general indication is that the information curves from the six selected test forms were essentially centered at the same point as the target curve, but with nominally larger variances.

Insert Figure 1 about here

Figure 2 presents 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 curves across content areas must take into account the scaling of the ordinate axes. For example, the AT Forms appear to demonstrate a greater lack of fit than the AAR Forms. However, if we consider the ordinate axes of the AT curves versus the AAR curves, it should be obvious that the real differences between the AT curves (2 items per subtest form) are actually as small or smaller than the differences between the AAR curves (14 items per subtest form).

Insert Figure 2 about here

In judging the actual degree of fit between curves, a more useful set of goodness-of-fit indices (beyond visual inspection) seems needed. Table 3 presents four such indices for the six AAP Math Forms fit to the Form 26A target information.

Insert Table 3 about here

The unweighted average absolute difference (|UAD|) represents the mean of the unsigned differences between the curves, as given by

$$|UAD| = \frac{\sum_{k=1}^{K} |T_k - \hat{T}_k|}{K} \qquad (14)$$

The unweighted root mean square (URMS) index represents the square root of the mean squared deviations between the fitted and target curves along the quadrature points. That is,

$$\mathrm{URMS} = \sqrt{\frac{\sum_{k=1}^{K} (T_k - \hat{T}_k)^2}{K}} \tag{15}$$

The weighted mean square (WMS) is similar to the URMS, but uses a normalized weighting of the standardized true scores, given each quadrature point, to essentially scale the information differences to the expected score density of the θ metric for the selected items. Therefore, the weighted mean square is given by

$$\mathrm{WMS} = \sum_{k=1}^{K} \phi(\zeta_k)\ (T_k - \hat{T}_k)^2 \tag{16}$$

where

$$\phi(\zeta_k) = \frac{e^{\frac{-\zeta^2_k}{2}}}{\sum_{k-1}^{K} e^{\frac{-\zeta^2_k}{2}}} \tag{17}$$

and

$$\zeta_k = \frac{\sum_{j-1}^{J} P_j(\theta_k) - \sum_{k-1}^{K} \sum_{j-1}^{J} P_j(\theta_k) / K}{\sqrt{\frac{K\sum_{k-1}^{K} \sum_{j-1}^{J} P_j(\theta_k) - [\sum_{k-1}^{K} \sum_{j-1}^{J} P_j(\theta_k)]^2}{K(K-1)}}} \tag{18}$$

given $P_j(\theta_k)$ as the probability of a correct response to item j, conditional upon θ_k. Finally, delta (Δ) is given as a squared difference weighted by the normalized information functions (densities) of the target function. Accordingly, delta becomes

$$\Delta = \sum_{k=1}^{K} \tau_k \, (T_k - \hat{T}_k)^2 \tag{19}$$

where

$$\tau_k = \frac{T_k}{\sum_{k-1}^{K} T_k} \tag{20}$$

By themselves, the four goodness-of-fit indices provided in the upper half of Table 3 imply both weighted and unweighted functions of various forms of the average unsigned differences between the Form 26A target curve and the selected test information curves (i.e. the curves for Forms A-F). However, to put these indices in a different perspective, we might consider these indices as proportions of an information function, conditional upon some value of θ. To do so merely requires dividing the value of the indice in Table 3 by the information function at some point along the θ metric (e.g. the mean information for the Form 26A target test of 21.67). For example, the |UAD|, URMS, WMS and Δ values (0.709, 0.849, 0.874 and 0.943) in the first row of Table 3 could be seen to represent proportional differences between the Form A curve and the target curve ranging from 3.28% to 4.36%, at the point of average test discrimination. These proportional differences, conditional upon the mean information in the Form 26A target curve, are provided in parentheses below each goodness-of-fit index in Table 3. The basic implication is that the fit between the information curves is actually far better than the indices in the upper half of Table 3 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 curve (e.g. 3.1% to 5.0% in terms of $|\text{UAD}|$) are essentially inconsequential.

As another method of assessing the goodness-of-fit, we might consider the relationship between the test information and the standard error of the latent abilities, θ, given by

$$\sigma_{e(\theta)} = \frac{1}{\sqrt{\sum_{j-1}^{J} I_j(\theta)}} \text{ .}$$

Using this relationship, it becomes possible to restate the goodness-of-fit statistics as weighted functions of the average unsigned differences between the standard errors conditional on θ. These standard error differences are provided in the lower half of Table 3.

The unweighted absolute average difference $\left(|\text{UAD}|_{SE(\theta)}\right)$ and the unweighted root mean square $(\text{URMS}_{SE(\theta)})$ of the standard errors obviously appear larger than the weighted mean square $\left(\text{WMS}_{SE(\theta)}\right)$ and delta $\left(\Delta_{SE(\theta)}\right)$. The reason has to do with the larger standard errors on θ at the asymptotes of the information curves. Because both the $|\text{UAD}|_{SE(\theta)}$ and $\text{RMS}_{SE(\theta)}$ indices treat all quadrature points of θ equally, both statistics essentially inflate the apparent unsigned average differences between the standard errors for the target versus fitted curves. $\text{RMS}_{SE(\theta)}$ further takes the square root which inflates the difference even more for values between 0 and 1. The $\left(\text{WMS}_{SE(\theta)}\right)$ and $\left(\Delta_{SE(\theta)}\right)$ indices, therefore, appear to be more meaningful in that both tend to limit the impact of standard error differences for θ values near the asymptotes. This is especially true if we consider that the seemingly largest differences between fitted forms and the target information functions occurred for Form F (referring to the upper half of Table 3). However, considering the weighted differences between the standard errors (lower half of Table 3), the differences are negligible.

Expected Score Differences

The final determinants of the adequacy and accuracy in fitting a target test using the 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 we consider the issue of parallelism among test forms to extend beyond our objective function (test information), then we must also consider what the score distributions of the fitted test forms will look like in comparison to the target test (AAP Math Form 26A, in this case).

Figure 3 presents the test characteristics curves (TCCs) for each of the six fitted test forms along with the TCC for Form 26A. These TCCs are defined by the sum of the conditional probabilities for all items in a test across the θ metric. That is,

$$T(\theta) = \sum_{j=1}^{J} P(\theta) \tag{21}$$

where $P_j(\theta)$ is the probability of a correct response to item j, conditioned upon θ (see equation [13]). T(θ) therefore defines the expectation of a random individual's true score on J items, given his/her ability level (Lord, 1980).

Insert Figure 3 about here

Quite clearly, 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-right scale. In the present study, predicted scores were obtained by assuming a (0,1) normal distribution on θ. Table 4 provides the means, standard deviations, 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 p-values and biserial correlations and their standard deviations are also shown in Table 4.

Insert Table 4 about here

Table 4 provides fairly clear evidence of parallelism among the six fitted forms and the target test, not only in terms of predicted means and standard deviations, but also skewness and kurtosis. In other words, the process of fitting the target information was sufficient to fit the expected and predicted score distributions for the present item pool. Finally, as suggested by the mean p-values and biserial correlations (and their standard deviations) the S_m and S_{rm} algorithms also seem to satisfy classical testing theory criteria for parallelism.

Microcomputer Timed Performance

ITEMSEL was run on a Compaq 386/33 microcomputer for the present study. As such, resulting performance indicators are perhaps optimistic ones for most microcomputer environments. Also, due to the interactive nature of the fitting process, user skill greatly enters into the assessment of timed performance. Nonetheless, several timing indices can be stated.

The entire process of constructing the six AAP forms, including all user inputs, fitting of subtargets and optimization of item content-designated subsets to the overall target information curve ranged from 15 to 20 minutes in multiple trials. This compares to informal estimates made by ACT test development staff of about 170 hours to accomplish the same task manually (Noble, 1990). Of course, the 170 hours would also include formulating additional constraints and making qualitative judgements about the constructed forms beyond the test information fit criteria.

In terms of more precise time estimates, the fitting of the six item subsets (six forms each) ranged from 1.2 to 10.9 seconds, depending upon the number of items. The process of choosing optimal subsets took 1.5 seconds of CPU time. Comparatively, fitting six forms of the overall Form 26A target curve (without content breakdowns) used 70.7 seconds of CPU time on the same Compaq 386/33 microcomputer. It should be noted, however, that these timing values also include the generation of graphics displays during all selection stages.

Discussion

The S_m and S_{rm} algorithms were introduced as viable methods for fitting test items to a target information curve. Both algorithms use the criterion of a moving average of the conditional distance to the target function, across quadrature points of θ. Items are then selected by use of a weighted composite score which assesses their fit to the criterion.

This approach appears to demonstrate three distinct benefits. First, the moving average criterion, as a form of an objective function, absorbs and redirects error in fit thus allowing for a non-iterative solution. The result is a reasonably fast method of fitting any target information curve. Second, the algorithms simultaneously consider all quadrature points which define the test information curves and upon which the information functions are conditional. That is, the entire information curve is always fit in the process of selecting items or items subsets. Finally, the algorithms can be conveniently extended for use with subtests/subtargets, item subsets and multiple test forms.

In general, ITEMSEL was able to produce six test forms which reasonably matched the Form 26A target test along multiple levels of 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 Mathematics test. Other criteria denoting the fit of the selected test forms to the target test (e.g., comparisons of the actual information curves) likewise demonstrated a strong association between forms.

The crucial point appears to be that ITEMSEL was able to successfully generate test forms with similar information curves. This was even shown to be the case when extending the notion of parallelism to expected score distributions and classical item parameters.

The process is, of course, far from perfect. Nonetheless, from an applied viewpoint: (a) the method is fast (which makes it feasible for microcomputer technology, even for large scale applications) and (b) it appears to be _at least_ as accurate as manual test construction methods given the constraints of this study. When implemented as part of an integrated software package such as ITEMSEL, these methods should readily complement the test construction process. This applied viewpoint defines the final intent behind the methods described in this paper.

Author Notes

[1] Partitioning the information CDF into equal areas essentially prioritizes the quadrature points of θ relative to the conditional information densities. Accordingly, the concentration and spread of θ corresponds closely to the actual distributional properties of the test information function.

REFERENCES

Ackerman, T. A. (April, 1989). An alternative methodology for creating parallel test forms use 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 Report 88-4, University of Twente, Netherlands.

Enschede: University of Twente, Department of Education, 1-10.

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: Lawrence Erlbaum Assoc.

Microsoft Corp. (1987). Quick Basic 4.0.

Nobel, C. (March, 1990). Personal communication.

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

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

van der Linden, W. J. & Boekkooi-Timminga, E. (1989). A maxmin model for test design with practical constraints. Psychometrika, 54(2), 237-247.

Wingersky, M. S., Barton, M. A., and Lord, F. M. (1982). LOGIST IV.

Table 1

Descriptive Summary of Fitted Item Parameters to Form 26A Target (40 Items)

	Means			Standard Deviations			Skewness			Kurtosis		
Form	a	b	c	a	b	c	a	b	c	a	b	c
AAP MATH 26A*	1.03	0.29	0.16	0.40	0.60	0.04	0.92	-0.62	0.03	0.19	-0.47	1.19
Form A	1.03	0.35	0.17	0.29	0.52	0.06	0.87	-0.20	0.52	0.96	-0.88	0.20
Form B	1.05	0.35	0.17	0.29	0.50	0.06	1.03	-0.31	2.17	1.68	-0.79	9.65
Form C	1.05	0.31	0.17	0.30	0.54	0.05	0.71	-0.12	0.25	1.09	-1.06	-0.25
Form D	1.05	0.32	0.16	0.29	0.55	0.06	1.46	-0.11	0.81	2.56	-0.66	1.50
Form E	1.04	0.31	0.17	0.28	0.50	0.06	1.25	-0.49	-0.32	1.88	-0.64	0.09
Form F	1.01	0.32	0.15	0.29	0.50	0.05	0.68	0.13	-0.27	0.88	-0.94	0.27

*Target set of items

Table 2

Means and Standard Deviations of IRT Parameters for 12 AAP Math Forms (Manually-Constructed)

(N = 40 Items)

Test Form	a	b	c
Form 24B	1.058 (0.296)	.309 (.661)	.160 (.084)
Form 25B	0.994 (0.247)	0.395 0.973)	0.159 (0.079)
Form 25C	1.078 (0.379)	0.359 (0.744)	0.157 (0.077)
Form 25D	1.068 (0.353)	0.321 (0.830)	0.142 (0.079)
Form 25E	1.057 (0.259)	0.307 (0.633)	0.128 (0.055)
Form 25F	0.950 (0.370)	0.385 (0.863)	0.152 0.062)
Form 26B	0.989 (0.358)	0.240 (0.875)	0.172 (0.046)
Form 26C	0.930 (0.365)	0.328 (0.876)	0.162 (0.034)
Form 26D	0.951 (0.427)	0.392 (1.283)	0.185 (0.026)
Form 26E	0.972 (0.297)	0.254 (0.777)	0.166 (0.048)
Form 26F	0.926 (0.365)	0.342 (0.953)	0.159 (0.034)
Form 27A	0.990 (0.394)	0.332 (0.868)	0.178 (0.046)

( ) = Std. Deviation

Table 3

Goodness-of-Fit Indices (to Form 26A Target)

	Information Function Indices			
Test Form	$\lvert UAD \rvert$	URMS	WMS	$\underline{\Delta}$
Form A	0.709	0.849	0.874	0.943
	(0.033)	(0.039)	(0.040)	(0.044)
Form B	0.681	0.788	0.637	0.644
	(0.031)	(0.036)	(0.029)	(0.030)
Form C	0.816	0.949	1.018	1.051
	(0.038)	(0.044)	(0.047)	(0.049)
Form D	0.733	0.889	0.733	0.688
	(0.034)	(0.041)	(0.034)	(0.032)
Form E	0.670	0.781	0.655	0.655
	(0.031)	(0.036)	(0.030)	(0.030)
Form F	1.078	1.276	1.885	1.944
	(0.050)	(0.059)	(0.087)	(0.090)

	$SE_{(\theta)}$ Indices			
Test Form	$\lvert UAD_{SE(\theta)} \rvert$	$URMS_{SE(\theta)}$	$WMS_{SE(\theta)}$	$\Delta_{SE(\theta)}$
Form A	0.040	0.159	0.006	0.0005
Form B	0.053	0.219	0.011	0.0010
Form C	0.039	0.146	0.004	0.0005
Form D	0.041	0.161	0.005	0.0006
Form E	0.039	0.150	0.005	0.0005
Form F	0.031	0.089	0.002	0.0003

( ) Proportion of mean information in the Form 26A target curve (21.67).

Table 4

Predicted Score Distributions for Six Fitted Test Forms and Target Form 26A

Test Form	$\bar{p}$	S_p	$\bar{r}_{bis}$	S_r	$\bar{X}$	S_x	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

*Target

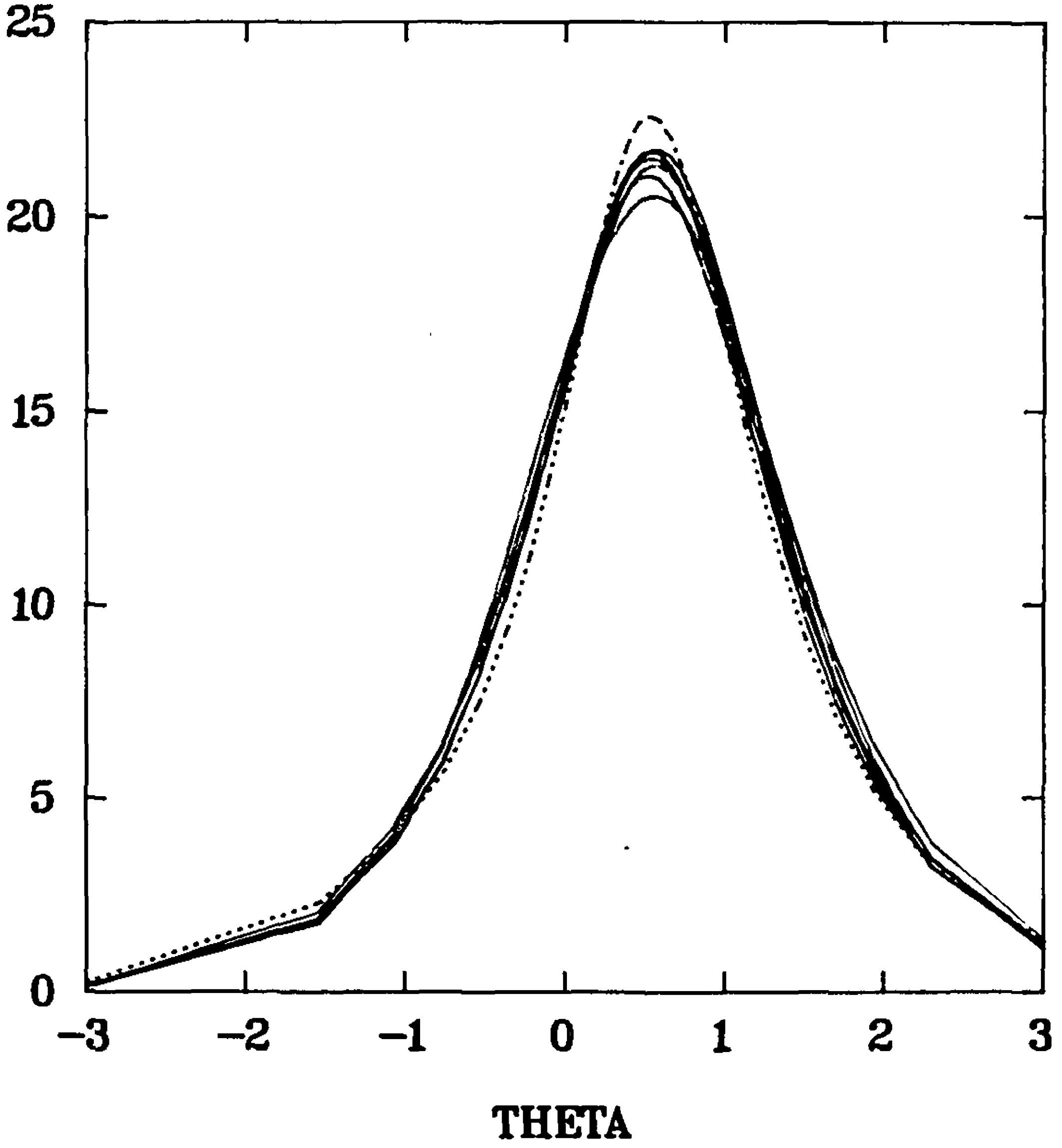
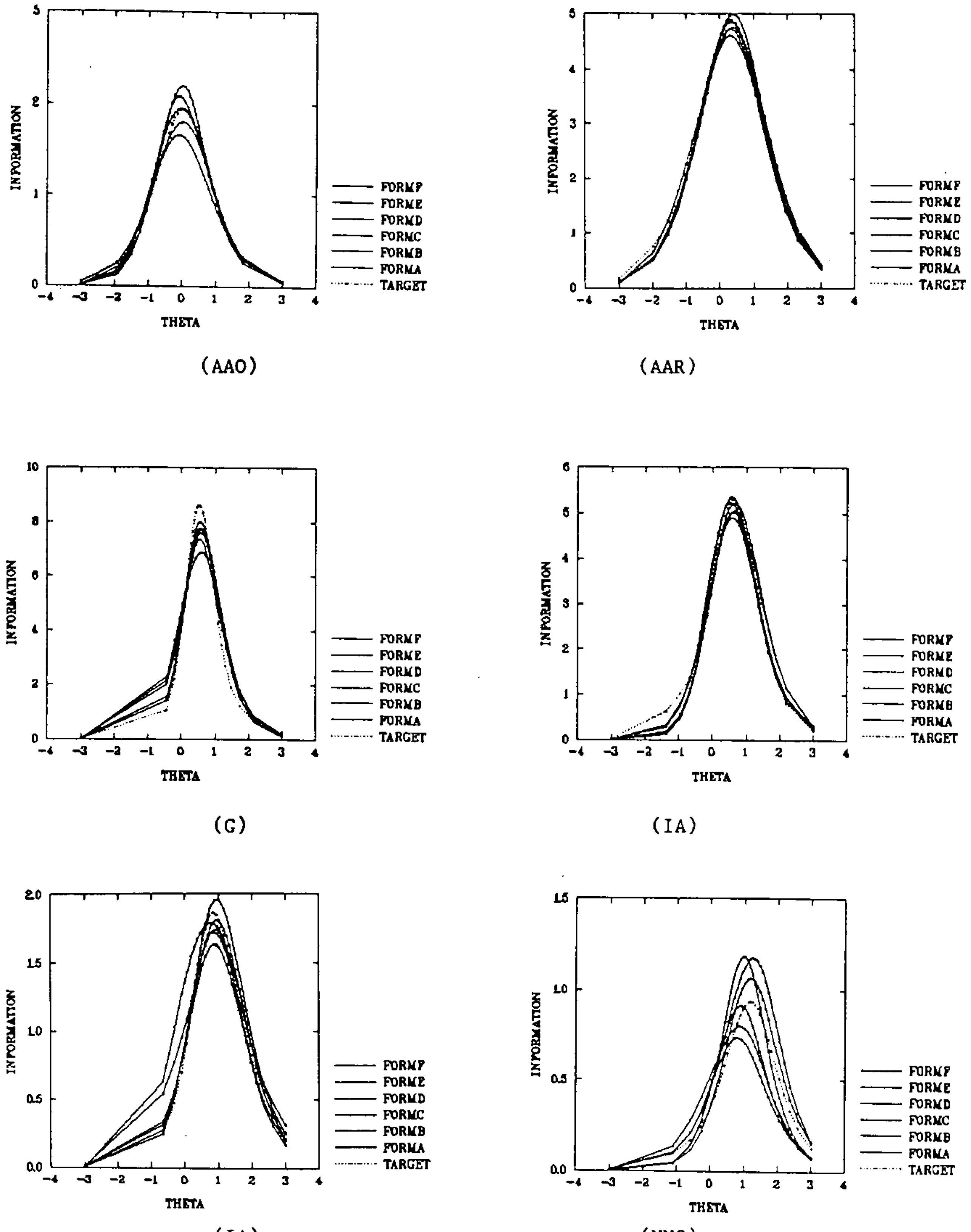


Figure 1. Test Information Curves for Six Forms of AAP Mathematics Test Fit to Form 26A Target Information Curve



(AAO)

(AAR)

(G)

(IA)

(IA)

(NNS)

Figure 2. Sub-content Fitting to Form 26A Content Items

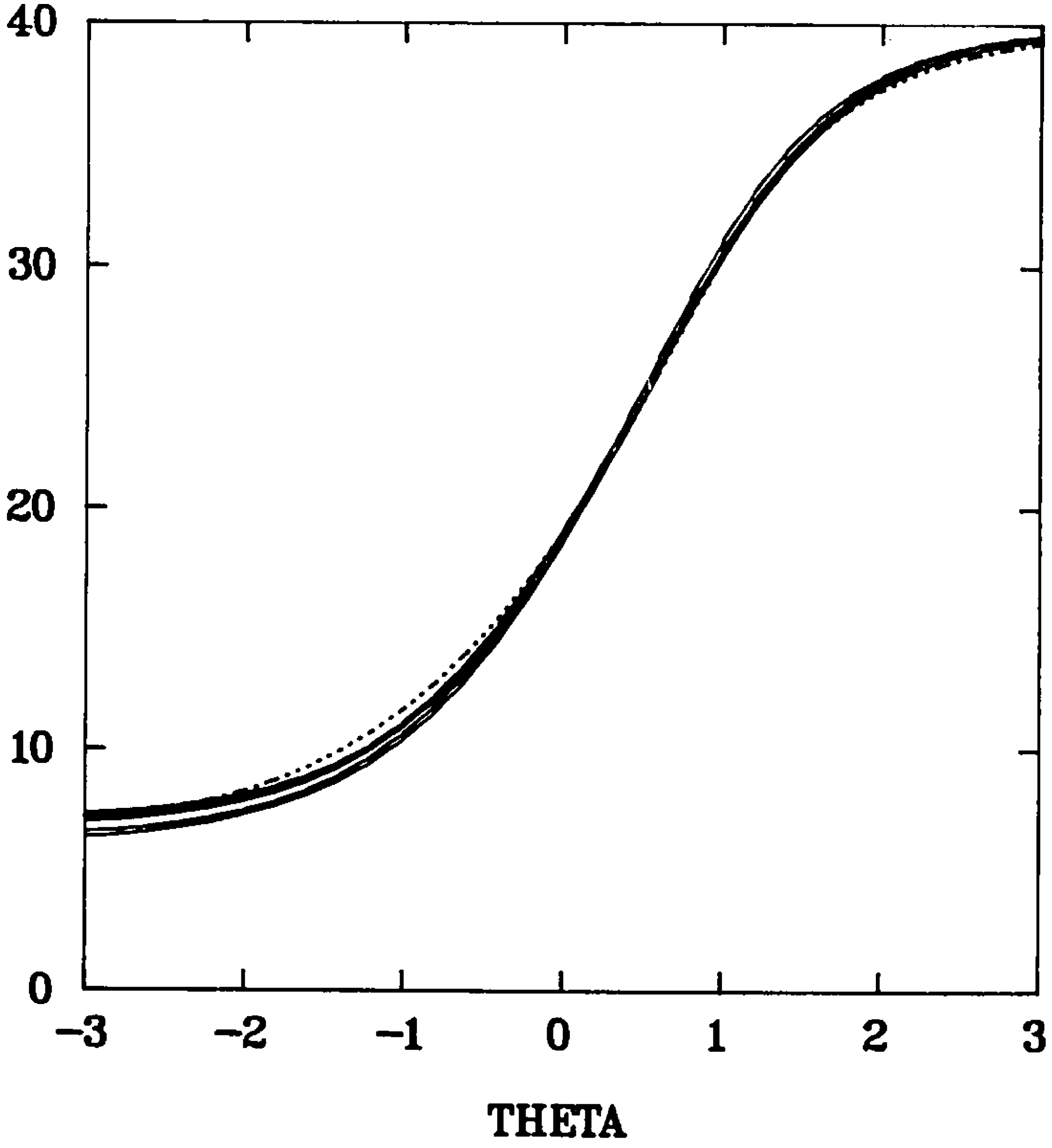


Figure 3. TCCs for the Target Test Form 26A and Six Test Forms Fitted by ITEMSEL