Test Construction by Means of Linear Programming

Dato N. M. de Gruijter
University of Leiden

The use of linear programming in the selection of test items entails setting a target information value for several ability levels, then constructing a test of minimum length that satisfies the constraints given by the target values. In the present paper the case of the uniform target is reconsidered. The dependency of item selection on item pool characteristics is demonstrated, and the relevance of uniform targets for test construction and the applicability of linear programming for test construction are discussed. *Index terms: item response theory, item selection, linear programming, test length.*

The relationship between item characteristics and test characteristics has long been an area of study. Since Birnbaum (1968), optimal item selection generally has been discussed within the framework of item response theory (IRT). In this context, the maximum amount of information at ability level θ is given by the test information, the sum of the item informations; the maximum is obtained under maximum likelihood estimation of θ. For each θ the target information, the minimum acceptable value for the test information, is specified.

Two cases can be discerned: mastery testing and conventional testing. In mastery testing the target information is high at only one θ level and item selection for minimum test length is relatively straightforward (Birnbaum, 1968; de Gruijter & Hambleton, 1983; Hambleton & de Gruijter, 1983).

With conventional tests the target information is essentially nonzero within a relatively large range of θ, and selecting the shortest test for which the test information at least matches the target information becomes a nontrivial task. Lord (1977) suggested a heuristic procedure to obtain a test of reasonable length. A major improvement was Theunissen's (1985) introduction of binary programming, a special case of linear programming (LP). In order to be able to apply binary programming to the problem of finding a test of minimum length, Theunissen had to replace the target information curve by target information values for a number of θ values.

Following a brief discussion of the LP approach, a result of Baker, Cohen, and Barmish (1988) with respect to a uniform target is replicated. It is demonstrated that this result depended on characteristics of the item pool. Finally, it is shown that a small deviation from a uniform target can give a considerable decrease in the number of items needed, and the reasonableness of a uniform target is discussed.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 2, June 1990, pp. 175–181

0146-6216/90/020175-07$1.60

The Linear Programming Approach

The test constructor specifies the target information $I(\theta)$ at m points θ_k ($k = 1, \ldots, m$) on the θ

scale. The function to be minimized is

$$n = \sum_{i=1}^{N} x_i \quad , \tag{1}$$

where N is the number of items in the item pool and $x_i = 1$ if the item is included in the test and 0 otherwise, subject to the constraints

$$\sum_{i}^{N} I_i(\theta_k)x_i \geq I(\theta_k) \qquad (k = 1, \ldots, m) \quad , \tag{2}$$

where $I_i(\theta_k)$ is the item information at θ_k. $I(\theta_k)$ is assumed to be accurately determined. Estimation errors or, worse, errors due to model misspecification (de Gruijter, 1986) are assumed to be negligible. With LP, which is less time-consuming than binary programming, individual x_i values are constrained to $0 \leq x_i \leq 1$. Hence fractional item contributions to the total test are possible in LP. However, the number of fractional values cannot exceed the number of constraints m (Dantzig, 1963). When a fractional contribution results, the corresponding item will be included in the test.

A Demonstration With a Uniform Target

In order to study item selection for a broad-range test (i.e., a test for measuring over a broad θ range), Baker et al. (1988) introduced uniform targets. Their choice is discussed critically below. One uniform target was defined in the range $-3 \leq \theta \leq 3$, with $I(\theta) = 16$; when θ is normally N(0,1) distributed, a uniform information curve with $I(\theta) = 16$ in the range $(-\infty,\infty)$ corresponds to $r_{\hat{\theta}\theta'} = .94$. In order to be able to use LP, Baker et al. specified 11 equally spaced points in the range $(-3,3)$. Finally, they generated a pool of 500 items conforming to the Rasch model by random sampling of item difficulty parameter values b from a N(0,1) distribution. They also generated item pools for the two- and three-parameter models; however, the present discussion considers only the Rasch model.

In the present study 500 b values were generated in the same manner and 11 θ levels—with $\theta_1 = -3$ and $\theta_{11} = 3$—were chosen for which the target information value was set equal to 16. Linear programming routine E04MBF from the NAG library (Numerical Algorithms Group, 1987) was used for solving the LP problem discussed above. The resulting minimum-length test consisted of 193 items, which exceeds the test length obtained by Baker et al. by only one item. Figure 1 gives the obtained information curve. The b distribution of the selected items is presented in Figure 2; this figure is comparable to Figure 2a in Baker et al.

On the basis of similar results for uniform targets and results for peaked targets, Baker et al. (1988) concluded that the LP algorithm—and for that matter, binary programming—attempts to cope with the "worst" feature of the target information curve. In the case of a uniform target, this consists of the

Figure 1
Obtained Information (Diamonds) and Target Information Level (Dotted Line) Based on a Pool With 500 Items

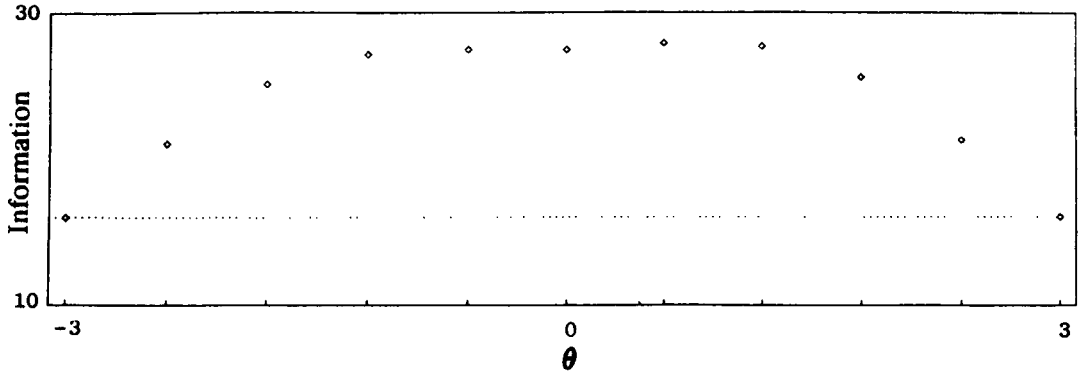
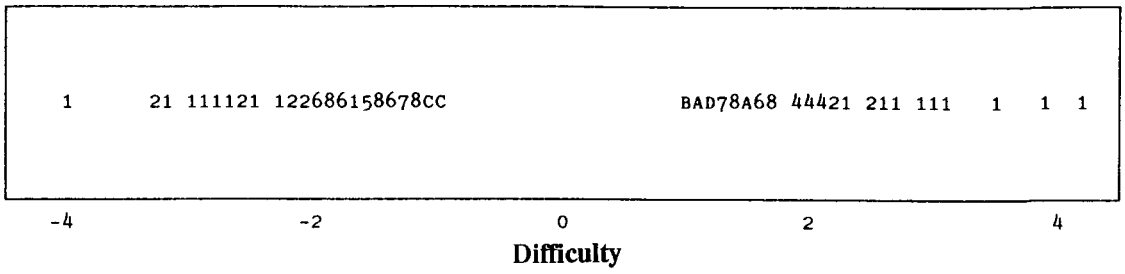


Figure 2
Distribution of Difficulty Parameters for the Selected Items Corresponding to Figure 1 ($A = 10$, $B = 11$, $C = 12$, $D = 13$)



extremes of the target, θ_1 and θ_m. This description is rather vague.

Figure 2 shows in more detail what has happened. Several items with $|b| > 3$ were selected. In theory there are better items. For example, item j with $b_j > \theta_m$ is dominated by all items i with $2\theta_m - b_j < b_i < b_j$: All of these items have higher item information values at the selected θ levels θ_k ($k = 1, \ldots, m$). Even so, items with extreme b values are selected for the simple reason that the item pool is depleted of more optimal items. The size of the item pool and the distribution of b values in the pool determine the final result. When the number of items in the pool increases, more items with adequate b value become available and the composition of the minimum-size LP test changes. This change also affects minimum test length: When more adequate items become available, minimum test length decreases.

Optimal Item Selection With No Hidden Restrictions

The next problem is the length and composition of an optimal test when availability of items does not constrain the LP solution. This can be addressed using a cluster-based approach suggested by Boekkooi-Timminga (1988). In order to reduce computing time, she defined $N^* < N$ intervals on the latent scale and set the item parameters equal to the midpoints of these intervals. With small intervals, the original distribution of b is adequately approximated.

With item clusters the function to be minimized is

$$n = \sum_{i=1}^{N^*} n_i \quad , \tag{3}$$

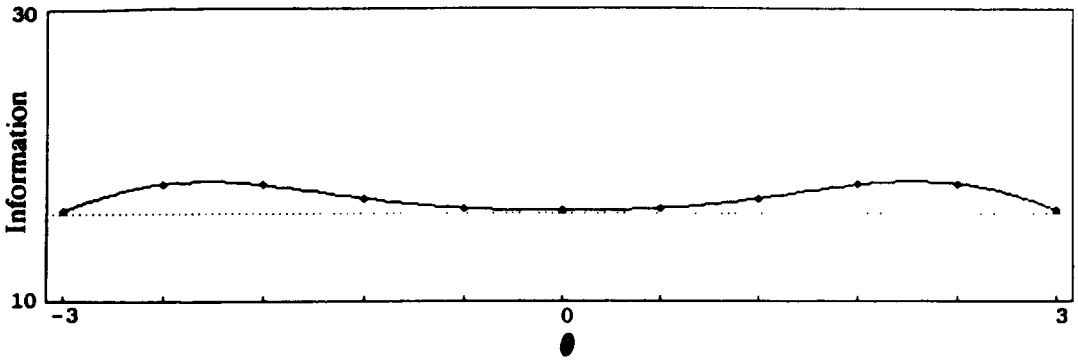
where n_i is constrained to positive integer values not exceeding the number of items in cluster i. The number of values to be determined is now only N^*.

For the present problem, $N^* = 121$ equally spaced values b_i ($i = 1, \ldots, N^*$) were specified in the relevant θ range $(-3,3)$. In order not to restrict the solution in some unknown way, a very large number was specified as the upper bound for n_i ($i = 1, \ldots, N^*$). The resulting n_i value is the LP solution corresponding to cluster i ($i = 1, \ldots, N^*$): n_i is the number of items needed from cluster i. The resulting values n_i can be fractional under LP. When only a few fractional values n_i are obtained, the next higher integer values can be chosen for the solution.

The LP solution for the uniform target obtained with the modified minimization problem on basis of $N^* = 121$ differs strongly from the solution depicted in Figures 1 and 2. Only three nonzero n_i values were obtained, for $b = -2.55$, $b = 0.0$, and $b = 2.55$. The respective numbers of items were 61, 32, and 61, giving a total test length of 154 items.

The difference with the Baker et al. results is the concentration of items with $b = 0$. The information curve is presented in Figure 3. It is quite possible that the information drops below 16 for θ values between two adjacent levels for which the information is specified. This would be the case

Figure 3
Obtained Information (Curved Line) and Target Information Level (Dotted Line) When the Item Pool Does Not Restrict the Solution



when 11 Guttman items with difficulties exactly equal to the 11 values θ_k had been selected. In order to verify whether this happened in the present case, information values were computed for intermediate θ values as well.

The uniform target curve is approximated reasonably well. The hill in Figure 1 has become a valley. The empirical information curve now approaches the target information at three values instead of two. This result is relevant insofar as it shows the effect of the hidden limitations in the Baker et al. item pool. Clearly it is difficult to generalize on the basis of their results.

Two conclusions can be drawn from the present results. First, as might be expected, the LP solution to the minimization problem strongly depends on the available items in the item pool. As pool size increases, the test length can decrease along with a shift in the item selection. In this respect a large item pool pays off. Second, the distribution of *b* values is relevant. The optimal solution of Figure 3 is not a realistic possibility, but the result shows that with the uniform target a pool with relatively many extreme *b* values is more efficient. This should be a consideration in item construction.

The Relevance of a Uniform Distribution

Even the optimal solution gives a relatively large minimum test length. In this respect the extremes of the uniform target still are the "worst" feature. This is easily demonstrated with an example. The target information value at θ_1 and θ_m was lowered to 12, and the restriction $m + 1$ was added:

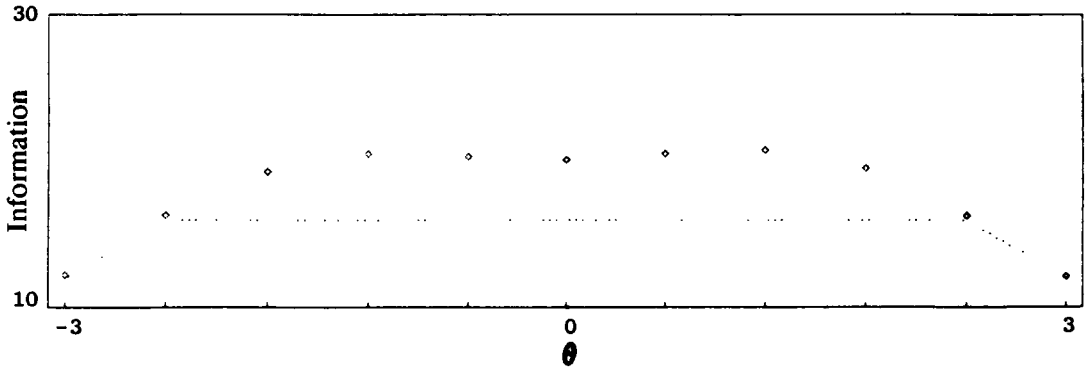
$$\sum_{i=1}^{N} \sum_{k=1}^{m} I(\theta_k) x_i \geq 196 \quad . \tag{4}$$

Hence the overall result of the LP solution (summing over all levels θ_k) should be higher than the implicit overall result in the first example, which equals 11 (the number of θ levels) × 16 (the target level), or 176. Nevertheless, in the pool of 500 items the solution gave a minimum test length of only 143 instead of 193 (or the minimum for the uniform target under optimal conditions, 154). A considerable reduction in the number of items could be obtained by less stringent demands at the extremes. The resulting information curve is given in Figure 4.

The resulting item selection is presented in Figure 5. The most conspicuous difference with respect to the selection in Figure 2 is the absence of very extreme *b* values: The modified target information draws less on the relatively scarce items with very high or low *b* values.

Now reconsider the connection between uniform targets and the corresponding item selections for broad-range tests. With respect to test composition, Gulliksen (1945) argued in favor of peaked tests, contrary to the practice common in those days. His

Figure 4
Obtained Information (Diamonds) With Lowered Target Information Levels at the Extremes (The Dotted Line Segments Represent the Target Information Level)



point was repeated by Cronbach and Warrington (1952). With the item discriminations generally found in practice, in IRT terms, peaked tests have high information levels for a very large range of θ. The test is accurate with respect to a particular population when the information peak coincides with the population mean.

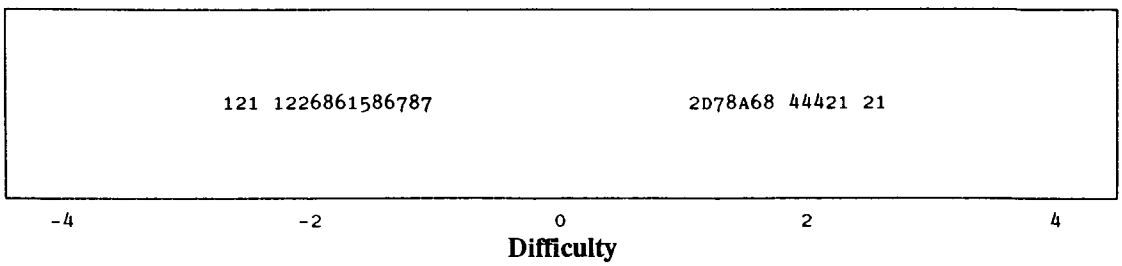
Lord (1985) provided a more precise account of the accuracy of tests in the context of IRT. He considered the problem of building an n-item test that minimizes expected loss over a single target population of examinees. From his results it can be concluded that when the loss structure is the same for all θ levels, expected loss is minimized if

$$I(\theta) = cg(\theta)^{1/2} \quad , \qquad (5)$$

where c is a constant and $g(\theta)$ denotes the frequency distribution (see Lord, 1985, Equation 10). Lord's approach to broad-range tests may lead to non-uniform targets. When the target θ distribution is normal, Equation 5 might be used to derive the target information curve instead of using a uniform target. A uniform target is relevant in this situation only when the target θ distribution is very flat.

Two additional analyses were done on basis of Equation 5 with normal θ distributions. In the first, a normal distribution with a mean of 0 and standard deviation of 1 was used in the derivation of the shape of the target information function. Minimization of Equation 3 without upper bounds on the values x_i led to an optimal test for which all items had $b = 0$. This analysis corresponds with previous

Figure 5
Distribution of Difficulty Parameters for the Selected Items Corresponding to Figure 4 ($A = 10$, $D = 13$)



suggestions by Gulliksen (1945) and Cronbach and Warrington (1952).

In the second analysis the standard deviation was set equal to 1.7, which corresponds to a standard deviation of 1 in the one-parameter normal ogive model; a peaked test was no longer optimal. As in the analysis for the uniform target information, three clusters of items were obtained. Cluster 1 contained items with $b = -2.05$ or -2.0, Cluster 2 contained items with $b = 0.0$, and Cluster 3 contained items with $b = 2.0$ or 2.05. The middle cluster contained more items than the other two clusters. Figure 6 gives the target information values and the obtained test information based on $I(\theta) = 25.0$ for $\theta = 0$.

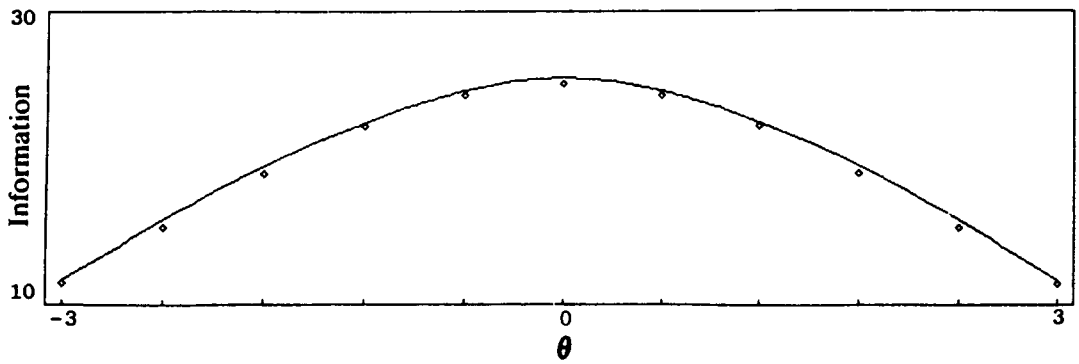
Discussion

The length and composition of an optimal test is highly dependent on the availability of enough items of an adequate difficulty level. Therefore, when several test forms are constructed sequentially from the same item pool, the length and composition of later tests can differ from those of the first test. Later tests become longer when the best items have been selected for previous test forms. A solution is to construct parallel tests simultaneously, as suggested by Boekkooi-Timminga (1988). This approach has limitations of its own, however: The number of test forms must be specified beforehand, and the possibility that the item pool expands over time cannot be taken into account. Such problems occur with all optimal item selection techniques, not just LP techniques.

When high discrimination is needed at only one θ level, there is no need for the introduction of a target information level; another more easily understood criterion is available (Hambleton & de Gruijter, 1983). In connection with broad-range tests, however, the specification of the target information is a point of concern. Although there is little doubt about the adequacy of a uniform target for broad-range testing based on tailoring of test items, the choice for a uniform target is less obvious in connection with the construction of a particular test form.

Work by Lord (1985) suggests another approach to test construction in which knowledge about an examinee population is relevant. This knowledge can be used in the specification of the target information. A well-considered choice of the target information is important because of its impact on the final solution for the optimal test. Uniform targets can lead to heterogeneous tests that are exceedingly long and result in high testing costs. In two cases Baker et al. (1988) depleted their item pool completely and still were unable to match or exceed the target information. A test constructor faced with such an outcome may well begin to question the correctness of the target specification. Uncertainty with respect to the target is best laid down formally in the test construction procedure.

Figure 6
Obtained Information (Solid Line) and 11 Target Values (Diamonds) Based on a Normal Target Population ($\sigma = 1.7$) in Case of Unlimited Item Resources



The procedure that led to the results depicted in Figure 4 gives an example. The overall target information was set at a relatively high level, but the target information at the extremes was lowered. This target structure did not exclude a solution in which all obtained information values at least matched the original uniform target level, but it allowed other solutions as well.

In summary, this study demonstrated that the conclusions of Baker et al. (1988) resulted in part from the limitations of their item pools. Further, as the technical problem of item selection is being solved with LP techniques, emphasis should shift to the specification of the target information function or, more generally, to the constraints for the minimization problem at hand. It can also be concluded that LP is useful not only in the stage of item selection, but also in earlier stages of test development. LP can be used to explore the optimal design of the item pool. When, for example, simulation results suggest that a heterogeneous item pool is most adequate in a particular application, item writers could be instructed to direct their efforts to writing items of varying difficulty. To obtain an acceptable test length, simulations might also be performed in order to obtain an indication of the minimum size of an item pool with a given distribution of item parameters.

References

Baker, F. B., Cohen, A. S., & Barmish, B. R. (1988). Item characteristics of tests constructed by linear programming. *Applied Psychological Measurement, 12,* 189–199.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Boekkooi-Timminga, E. (1988). *A cluster-based method for test construction* (Research Report 88-3). Enschede, The Netherlands: University of Twente.

Cronbach, L. J., & Warrington, W. G. (1952). Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika, 17,* 127–147.

Dantzig, G. B. (1963). *Linear programming and extensions*. Princeton NJ: Princeton University Press.

de Gruijter, D. N. M. (1986). Small *N* does not always justify Rasch model. *Applied Psychological Measurement, 10,* 187–194.

de Gruijter, D. N. M., & Hambleton, R. K. (1983). Using item response models in criterion-referenced test item selection. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

Gulliksen, H. (1945). The relation of item difficulty and interitem correlation to test variance and reliability. *Psychometrika, 10,* 79–91.

Hambleton, R. K., & de Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement, 20,* 355–367.

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

Lord, F. M. (1985). Estimating the imputed social cost of errors of measurement. *Psychometrika, 50,* 57–68.

Numerical Algorithms Group Inc. (1987). *Fortran Library Manual, Mark 12, Vol. 3*. Downers Grove IL: Author.

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

Author's Address

Send requests for reprints or further information to Dato N. M. de Gruijter, Educational Research Center, University of Leiden, Boerhaavelaan 2, 2334 EN Leiden, The Netherlands.