A Zero-One Programming Approach to Gulliksen's Matched Random Subtests Method

Wim J. van der Linden and Ellen Boekkooi-Timminga
University of Twente

Gulliksen's matched random subtests method is a graphical method to split a test into parallel test halves. The method has practical relevance because it maximizes coefficient α as a lower bound to the classical test reliability coefficient. In this paper the same problem is formulated as a zero-one programming problem, the advantage being that it can be solved by computer algorithms that already exist. It is shown how the procedure can be generalized to split tests of any length. The paper concludes with an empirical example comparing Gulliksen's original hand-method with the zero-one programming version. *Index terms: Classical test theory, Gulliksen's matched random subtests method, Item matching, Linear programming, Parallel tests, Test reliability, Zero-one programming.*

In order to estimate the classical coefficient of test reliability, parallel measurements are needed. Methods proposed to meet this requirement in practice include retesting the same examinees with the same test after some time has elapsed, or carefully constructing a parallel test and testing the same examinees with both instruments.

As is known from practical experience, however, these methods do not work well. The main objection against the test-retest method is that replicate test administrations are impossible with live examinees who may exhibit all kinds of interfering processes, such as remembering earlier administrations, learning and forgetting between administrations, or being less than optimally motivated to participate in another administration. The parallel-forms method, on the other hand, constitutes a dilemma. It assumes that it is possible to construct two *different* tests with exactly the *same* measurement properties. Practical experience shows that this ideal may be attained to some extent but is never realized exactly.

As a possible way out of this fundamental problem, Kuder and Richardson (1937) proposed their formulas 20 and 21 which can be estimated using (dichotomous) item and test scores from a single administration. A generalization of these formulas to non-dichotomous items or test components of any length is known as Cronbach's (1951) coefficient α:

$$\alpha \equiv \frac{n}{n-1}\left[1-\frac{\sum\sigma^2(Y_g)}{\sigma_X^2}\right] \qquad (\sigma_X^2 > 0,\ n > 1) \quad , \tag{1}$$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 2, June 1988, pp. 201–209

0146-6216/88/020201-09$1.70

where $\sigma^2(Y_g)$ is the variance of the scores Y_g on test component g,
σ_X^2 is the variance of the score X, and
n is the number of components.

The usual choices of test components in this internal-consistency method are the individual test items or test halves. Estimates of the test reliability based on the latter are known as split-half estimates. A generalization of Equation 1 to any split was introduced by Raju (1977) and is known as coefficient β_k.

Analysis of the relationship of Equation 1 to the definition of the reliability coefficient reveals that they are equal to each other only if the test components are essentially τ-equivalent; otherwise Equation 1 is a lower bound to the test reliability (e.g., Lord & Novick, 1968, pp. 87–95). Although this requirement is less restrictive than the one of parallel measurements, it seems to give rise to the same practical problems as for the test-retest and parallel-forms methods. However, there is a possibility of optimization that the latter methods do not possess. Because Equation 1 is a lower bound to the reliability for *any* split of the test into components, and these bounds are not necessarily equal, the split with the greatest lower bound can be used as the basis for estimation of the reliability coefficient. It is for this reason that the internal-consistency method has not only a practical but also some theoretical appeal.

Gulliksen (1950/1987) proposed a method for splitting tests optimally into halves, now known as the matched random subtests method. The method is the only one available for this important purpose and is described in most textbooks on test theory (e.g., Allen & Yen, 1979, pp. 78–83). Despite this, it has not been implemented in standard computer packages for test analysis and is rarely used on a routine basis; the likely reason is that the method is graphic and must be performed by hand. It is the purpose of this paper to present a version of Gulliksen's method that is derived from zero-one programming. Algorithms for this method exist and are amply available in computer code. Properties of this version of Gulliksen's method are explored using empirical test data.

Gulliksen's Matched Random Subtests Method

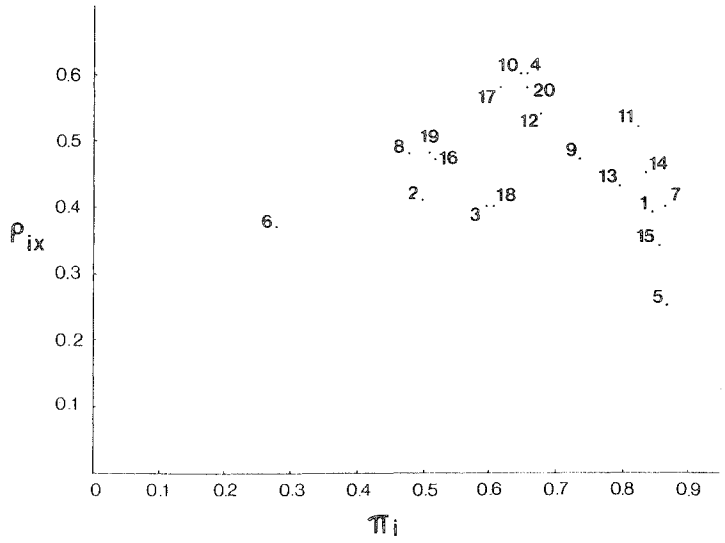
Gulliksen's method is usually formulated for dichotomous item scores but can easily be generalized to other situations. For dichotomous item scores, the method involves two parameters for each item: difficulty and discrimination. Let π_i and ρ_{iX} denote the classical definitions of these parameters. Then the former is the expected item score and the latter is the point-biserial correlation between the item and test score. Each item is plotted on a graph with its values for the two parameters as coordinates. Next, pairs of items are formed, the criterion being that each pair should have points on the graph as close to each other as possible. Test halves are obtained by assigning one randomly chosen item from each pair to one test half and the remaining items to the other.

Figure 1 shows a typical Gulliksen plot. The points are estimates for a 20-item version of a mathematics achievement test used in the Second Mathematics Study of the International Association for the Evaluation of Education based on a Dutch sample of 5,418 examinees. The same data are used in the empirical example below. Note that some pairs in Figure 1 are obvious. Others, however, are not. Item 16, for instance, could be paired with item 19 but this choice has consequences for the possibilities of item 8; the choice for this item, in turn, restricts the possibilities for item 2, and so on. In fact, it is the absence of a clear-cut criterion for coping with such dependencies that may make the method difficult to use for larger sets of items.

Let Y_g in Equation 1 represent the observed score on test half g which consists of n_g items ($g = 1, 2$). A well-known result from classical test theory is that, for dichotomous items, the expected values and variances of Y_g can be written as functions of π_i and ρ_{iY_g} only. Assuming $\rho_{iY_g} = \rho_{iX}$ for $g = 1, 2$, as is implicitly done in the Gulliksen method, the expressions are

$$\mu_{Y_g} = \sum_{i=1}^{n_g} \pi_i \qquad (2)$$

Figure 1
The Gulliksen Plot for a 20-Item Test



and

$$\sigma^2_{Y_g} = \left[\sum_{i=1}^{n_g} \pi_i(1-\pi_i)\rho_{iX}\right]^2 . \tag{3}$$

Gulliksen's method is motivated by the fact that pairwise matching of the items on π_i ensures that μ_{Y_1} and μ_{Y_2} are approximately equal. Hence, a necessary condition for the two halves to have the same true scores is met. As matching on ρ_{iX} also ensures approximately equal values of Equation 3 for $g = 1, 2$, the two halves may have equal error variances and meet the requirements of parallel measurements.

As already mentioned, Gulliksen's method is graphic. It supposes the presence of a judge inspecting the graph and matching the items in pairs. It is not an algorithm in the sense that all of its rules can be written in computer code. As illustrated earlier, its criterion for pairing the items is not unequivocal. Therefore, situations may arise where the judge does not know with certainty which of the possible pairs to select. Also, the random assignment of items from pairs to test halves may be suboptimal. In particular, when the items within pairs are not close to each other, there is a non-negligible probability that random assignment will result in test halves being less parallel than necessary.

Another desirable improvement on the method would be an algorithm equally well applicable to splits into other components than test halves. Splits of tests into thirds or quarters, for instance, require the division of the plots into triples or quadruples of items. It is unlikely that this can be done satisfactorily for larger tests merely by inspecting plots. On the other hand, such splits also yield values for Equation 1 that are lower bounds to the reliability coefficient, and it seems unwise to confine the search of the greatest lower bound only to the subset of splits into test halves.

Like any other method of item selection, the Gulliksen method poses the danger of capitalizing on chance if it is used with sample statistics instead of parameters. For this reason, it can only be recommended as a large-sample solution to the problem of splitting a test into parallel halves. The same holds if the zero-one programming formulation of Gulliksen's method given below is used with statistics instead of parameters.

A Zero-One Programming Version of Gulliksen's Method

Gulliksen's method consists of two steps—pairing the items and assigning items from pairs to test halves. Both tasks can be performed using techniques from zero-one programming. Interest in the application of zero-one programming techniques to problems in test theory originated in a recent paper by Theunissen (1985), who applied them to solve the problem of automated test design in item response theory. This problem is pursued further in Theunissen and Verstralen (1986) and van der Linden and Boekkooi-Timminga (in press), whereas Boekkooi-Timminga (1986, 1987) provided extensions to the problems of simultaneous test design and the design of parallel tests in item response theory. The techniques used below have a close relationship to the ones in the two Boekkooi-Timminga studies, but are applied here in the context of classical test theory; the minimax approach in van der Linden and Boekkooi-Timminga (in press) is also used.

Pairwise Item Matching

In Gulliksen's method the items are paired on inspection. It is suggested that this approach be replaced by the following unequivocal criterion. In the graph the Euclidean distance

$$\delta_{ij} = [(\pi_i - \pi_j)^2 + (\rho_{iX} - \rho_{jX})^2]^{1/2} \tag{4}$$

between the points i and j ($i \neq j$) is considered. It is proposed to pair the items such that the sum of the within-pair distances is minimal. In the following, as is necessary in the Gulliksen method, n is assumed to be an even number. [If n is odd, one item must be deleted and a Spearman-Brown correction with factor $n/(n-1)$ should be applied to the eventual reliability estimate.] Let x_{ij} be a binary decision variable denoting whether i and j are a pair. That is,

$$x_{ij} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ are a pair} \\ 0 \text{ otherwise} \end{cases} \qquad (i < j) \quad .$$

The problem is to decide on the $n(n-1)/2$ values of x_{ij} such that the criterion of a minimal sum of distances is met. Now the product $\delta_{ij}x_{ij}$ is equal to the distance between i and j if they are a pair, and to 0 otherwise. The problem is thus to minimize the sum of these products subject to the constraints that each item must be a member of exactly one pair. In the usual zero-one programming format the problem is as follows: Minimize

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta_{ij} x_{ij} \tag{5}$$

subject to

$$\sum_{i=1}^{j-1} x_{ij} + \sum_{i=j+1}^{n} x_{ji} = 1 \qquad (j = 1, \ldots, n) \tag{6}$$

$$x_{ij} \in \{0, 1\} \qquad (i = 1, \ldots, n-1; j = i+1, \ldots, n) \quad , \tag{7}$$

where for notational convenience the sums in Equation 6 are equal to 0 if the upper bound to the index is smaller than the lower bound, or conversely. The objective function in Equation 5 is defined as the minimization of the sum of all within-pair distances. The constraints in Equation 6 guarantee that for each item the decision variables x_{ij} ($i < j$) take the value 1 exactly once, which means that each item arrives in exactly one pair. In Equation 7 the decision variables are constrained to be binary.

The problem in Equations 5 through 7 is a standard zero-one programming problem that is found in textbooks on linear programming (e.g., Wagner, 1975, chap. 13). Algorithms to solve the problem

can also be found in Wagner and have been implemented in various computer programs. In the empirical example below, the program LANDO, which is based on the branch-and-bound method of Land and Doig (1960), was used. The output of the program is the $n(n-1)/2$ values of the decision variables x_{ij}, with $n/2$ values equal to 1 and the remaining values equal to 0.

Assigning Items to Components

The optimization procedure could stop here to randomly assign items from pairs to test halves, as is done in the Gulliksen method. However, it is also possible to match the test halves further, for instance, on their average scores or variances. In both cases the problem is again one of zero-one programming. If the latter option is chosen, the problem is to match the test halves on their sums of the terms $\pi_i(1-\pi_i)\rho_{iX}$ in Equation 3. Because, by definition, there are only two test halves, matching the two sums is equivalent to minimizing the sum with the larger value. Formulating the problem using this minimax criterion has the advantage that it can easily be generalized to other splits than test halves. This generalization will be shown below.

The output of the previous problem is a set of $n/2$ pairs. Let (p,q) denote the pth item $(p = 1, 2)$ in the qth pair $(q = 1, \ldots, n/2)$ and define a binary decision variable x_{pqr} $(r = 1, 2)$ as

$$x_{pqr} = \begin{cases} 1 \text{ if item } (p,q) \text{ is assigned to test half } r \\ 0 \text{ otherwise.} \end{cases}$$

Finally, let z be an arbitrary upper bound to the sums of $\pi_i(1-\pi_i)\rho_{iX}$ in the test halves. Then the assignment problem can be formulated as: Minimize z subject to

$$\sum_{p=1}^{2}\sum_{q=1}^{n/2}\pi_{pq}(1-\pi_{pq})\rho_{pq}x_{pqr} - z \leq 0 \qquad (r = 1, 2) \quad , \tag{8}$$

$$\sum_{p=1}^{2}x_{pqr} = 1 \qquad (r = 1, 2; q = 1, \ldots, n/2) \quad , \tag{9}$$

$$\sum_{r=1}^{2}x_{1qr} = 1 \qquad (q = 1, \ldots, n/2) \quad , \tag{10}$$

and

$$x_{pqr} \in \{0, 1\} \qquad (p = 1, 2; q = 1, \ldots, n/2) \quad , \tag{11}$$

where π_{pq} and ρ_{pq} are the item difficulty and discrimination indices for item (p,q). The constraints in Equation 8 ensure that the standard deviations of the two test halves are not larger than the minimized upper bound z. The constraint in Equation 9 requires that the items in a pair are assigned to different test halves each consisting of $n/2$ items; Equation 10 requires that each item is assigned exactly once. The constraints in Equations 9 and 10 could be simplified by replacing x_{pqr} with a variable x_{pq}, equal to 1 if (p,q) must be assigned to the first test half and equal to 0 otherwise, but then the generalization to splits other than test halves to be presented below is not so obvious.

The same analysis could be done with π_{pq} as coefficients in Equation 8 matching the test halves on their average scores, with weighted combinations $c\pi_{pq} + (1-c)\pi_{pq}(1-\pi_{pq})\rho_{pq}$ $(0 \leq c \leq 1)$ as coefficients, or with inequality constraints on the averages (variances) added to the model matching the test halves on their variances (averages). All of these options are attributable to the fact that the underlying problem of matching test halves on parallelness is one of multiple-objective decision making. This wealth of choices need not be bothersome, however, because the previous pairing of the items already ensures a high match of the test halves on both their averages and variances before they enter this stage of

optimization. In the empirical example below, weighted coefficients with $c = .5$ are used. This choice is in the same spirit as the first stage in Gulliksen's method, where in Equation 4 π_i and ρ_{iX} are also weighted equally.

Optimization Without Item Matching

In Gulliksen's method, two different steps are involved: First, pairs of matched items are found, and then items are assigned to test halves. It would seem that the first step is redundant if the model for the second step could be relaxed such that the items are assigned from the full set of n items instead of from $n/2$ pairs. The following model provides this relaxation: Minimize z subject to

$$\sum_{i=1}^{n}[c\pi_i+(1-c)\pi_i(1-\pi_i)\rho_{iX}]x_{ir}-z \leq 0 \qquad (r = 1, 2) \quad , \tag{12}$$

$$\sum_{i=1}^{2}x_{ir} = 1 \qquad (i = 1, \ldots, n) \quad , \tag{13}$$

$$\sum_{i=1}^{n}x_{i1} = n/2 \quad , \tag{14}$$

and

$$x_{ir} \in \{0, 1\} \qquad (i = 1, \ldots, n; r = 1, 2) \quad . \tag{15}$$

The difference between Equations 8 and 12 is that the latter has a weighted contribution of the items to the test-half averages and variances as a coefficient of the decision variables. The new constraint in Equation 14 requires that each test half be composed of exactly $n/2$ items.

At first glance, models such as the one specified in Equations 12 through 15 have greater capacity for optimization than the one in Equations 5 through 11. However, their use is not recommended; in general they produce less satisfactory results, because in each of the test halves compensation of the item properties is possible. For example, the result can be one test half with all items of moderate difficulty and another with items considerably varying in difficulty. As will be shown in an empirical example below, such test halves do not necessarily produce optimal reliability estimates. It is the possibility of compensating item properties that must have brought Gulliksen to the idea of previous item matching.

Generalization to Other Splits

Triples of Items

It is assumed that n is a multiple of 3. Then the within-triple "distance" is defined as $\delta_{ijk} = \delta_{ij} + \delta_{ik} + \delta_{jk}$ for all triples (i,j,k) $(i \neq j, j \neq k, i \neq k)$; the decision variable x_{ijk} is equal to 1 only if i, j, and k are in the same triple, and is equal to 0 otherwise $(i < j < k)$.

The problem is now: Minimize

$$\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n}\delta_{ijk}x_{ijk} \tag{16}$$

subject to

$$\sum_{j=2}^{k-1}\sum_{i=1}^{j-1}x_{ijk}+\sum_{i=1}^{k-1}\sum_{j=k+1}^{n}x_{ikj}+\sum_{j=k+2}^{n}\sum_{i=k+1}^{j-1}x_{kij} = 1 \qquad (k = 1, \ldots, n) \tag{17}$$

and

$$x_{ijk} \in \{0, 1\} \qquad (i = 1, \ldots, n-2; j = i+1, \ldots, n-1; k = j+1, \ldots, n) \quad , \tag{18}$$

where, for notational convenience, undefined sums in Equation 17 are again set equal to 0. The values in the upper and lower bounds in Equation 17 follow from the requirement that x_{ijk} be defined for $i < j < k$ only.

Assigning Items to Components

If in Equations 8 through 11 the indices are $p = 1, 2, 3$, $q = 1, \ldots, n/3$, and $r = 1, 2, 3$, the model assigns items from triples to test components of size $n/3$. This immediately suggests how the model can be generalized to splits into test components of any length.

An Empirical Example

In order to illustrate the procedures, the algorithm by Land and Doig (1960), as implemented in the program LANDO, was used together with the item data in Figure 1. The item difficulties and item-test correlations were estimated from a sample of 5,418 examinees, which is large enough to prevent capitalizing on chance in the Gulliksen method. The estimates are presented in Table 1.

As was clear from the bivariate distribution of the estimates in Figure 1, it is not immediately obvious how all of these items should be paired by hand. Table 2 gives the optimal item pairs following Equations 5 through 7. The results of the assignment of the items to test halves according to the optimization model in Equations 8 through 11 are presented in Table 2 by underscoring the items in the same test half. In making these assignments, the equally weighted sums of Equations 2 and 3 were used as coefficients in Equation 8.

Next, in order to compare empirical results of Gulliksen's method with the model in Equations 5 through 11, data from two test administrations were studied in more detail. One test was a Physics test for which the responses of a sample of 5,165 examinees to 20 items were available. In addition, the responses of a sample of 5,000 examinees to a 20-item Commercial Practice test were used. Both samples of examinees were large enough to prevent the conclusions from being dependent on sampling fluctuation. For each test a Gulliksen plot was prepared, and four persons, all trained in test theory and construction, were asked to perform Gulliksen's method by hand. In addition, the model in Equations 5 through 11 was used to split the tests into halves. In doing so, the coefficients $c\pi_{pq} + (1 - c)\pi_{pq}(1 - \pi_{pq})\rho_{pq}$ with $c = .50$ were used. However, later studies with $c = .10$, .25, .75, and .90 yielded exactly the same splits.

Table 1
Difficulty (π_i) and Discrimination (ρ_{iX}) Values for the Items in the 20-Item Test

Item	π_i	ρ_{iX}	Item	π_i	ρ_{iX}
1	.85	.39	11	.83	.52
2	.50	.41	12	.68	.54
3	.60	.40	13	.80	.43
4	.66	.60	14	.84	.45
5	.87	.25	15	.86	.34
6	.28	.37	16	.52	.47
7	.87	.40	17	.62	.58
8	.48	.48	18	.61	.40
9	.74	.47	19	.51	.48
10	.65	.60	20	.66	.58

Table 2
Optimal Item Pairs
and Test Halves
(Underlined Item Numbers
Were in the Same Test Half)

Matched Pairs (i,j)	$\lvert\pi_i-\pi_j\rvert$	$\lvert\rho_{iX}-\rho_{jX}\rvert$
1 - 7	.02	.01
2 - 6	.22	.04
3 - 18	.01	.00
4 - 10	.01	.00
5 - 15	.01	.09
8 - 19	.03	.00
9 - 13	.06	.04
11 - 14	.01	.07
12 - 20	.02	.04
16 - 17	.10	.11

The results are given in Table 3. All reliability estimates were calculated as correlations between test halves corrected for test lengthening by the Spearman-Brown formula. For the Physics test, the test constructors implementing Gulliksen's method and the model in Equations 5 through 11 performed equally well. Apparently, there was no space for further optimization to be used by the model. Although the differences were small, the results for the Commercial Practice test showed a different picture: The model produced an optimal lower bound to the test reliability, but some of the test constructors implementing Gulliksen's method were not able to do so. As a benchmark, the values of α are given. For both tests the model produced slightly larger reliability estimates. The same was the case for the split-half (first-second half) and odd-even methods.

For the same datasets the model in Equations 12 through 15 was used to split the tests into halves. The reliability estimates it produced were .77 (Physics test) and .61 (Commercial Practice test), showing that the model does not necessarily give optimal reliability estimates.

Conclusions

The main conclusion from the examples is that whenever there is space for optimization, the model exploits this and produces an estimate of a larger lower bound to the test reliability. Whether in practice there is more space than for the tests in Table 3, which apparently were rather homogeneous and yielded

Table 3
Results from Empirical Comparisons
Between Reliability Estimation Methods

Test	Split Half	Odd-Even	Model 4-11	Test Constructor 1	2	3	4
Physics	.75	.79	.79	.79	.79	.79	.79
Commercial Practice	.59	.59	.61	.61	.60	.56	.61

Note. For the Physics and Commercial Practice tests α=.78 and .57, respectively.

small differences between reliability estimates based on different splits, is an empirical matter. The attractiveness of the model in Equations 5 through 11 is that it *automatically* selects the optimal split with certainty, and that no hand-work is necessary. In addition, the same zero-one programming model can be used in any other situation where classically parallel tests are needed, such as in pretest-posttest designs in educational research or in situations where a test security problem exists.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey CA: Brooks/Cole.

Boekkooi-Timminga, E. (1986). *Algorithms for the construction of parallel tests by zero-one programming*. Enschede, The Netherlands: Department of Education, University of Twente.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. *Methodika, 1*, 101–112.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley. [Hillsdale NJ: Erlbaum, 1987.]

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Land, A. H., & Doig, A. (1960). An automatic method of solving discrete programming problems. *Econometrika, 28*, 497–520.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42*, 549–565.

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

Theunissen, T. J. J. M., & Verstralen, H. H. F. M. (1986). Algorithmen voor het samenstellen van toetsen [Algorithms for test construction]. In W. J. van der Linden (Ed.), *Moderne methoden voor toetsconstructie en -gebruik*. Lisse, The Netherlands: Swets & Zeitlinger.

van der Linden, W. J., & Boekkooi-Timminga, E. (in press). A maximin model for test design with practical constraints. *Psychometrika*.

Wagner, H. M. (1975). *Principles of operations research*. London: Prentice/Hall International.

Acknowledgments

This research was supported in part by a grant from the Dutch Organization for Pure Research (ZWO) through the Foundation for Psychological and Psychonomic Research in the Netherlands (PSYCHON).

Author's Address

Send requests for reprints or further information to Wim J. van der Linden, Faculteit der Toegepaste Onderwijskunde, Universiteit Twente, Postbus 217, 7500 AE Enschede, The Netherlands.