Maximizing the Validity of a Test as a Function of Subtest Lengths for a Fixed Total Testing Time: A Comparison Between Two Methods

Tami Kennet-Cohen, Shmuel Bronner and Yoav Cohen

Paper presented at the annual meeting of the National Council
on Measurement in Education, San Francisco, 2006

Abstract

Two methods proposed for determining the lengths of the subtests of a test with a fixed total testing time, so as to maximize the predictive validity of the test, were compared. In the *search method* (Kennet-Cohen, Bronner, & Cohen, 2003) a search for the optimal allocation of the total testing time among the subtests is conducted by a repetitive process of transferring testing time from one subtest to another and calculating the predictive validity that would be obtained. In the *analytic method* (Jackson & Novick, 1970), a formal solution is offered. This solution is valid and unique whenever it specifies nonnegative times for all subtests. A step-down procedure is suggested for cases in which some of the testing times are zero. Both methods were applied to the Psychometric Entrance Test, using data obtained from 4,321 first-year students in Israeli universities. Not only was the *search method* validated by the *analytic method*, it also overcame some of its limitations.
Two appendices are included in the paper. Appendix 1 presents a comparison between the *search method* and the *regression-weights* approach for maximizing validity. Appendix 2 explains and discusses a correction used in the calculation of the estimates of the subtest reliabilities.

Objective of the Inquiry

In a previous work (Kennet-Cohen, Bronner, & Cohen, 2003), a *search method* was proposed for determining the lengths of the subtests of a test so as to maximize the correlation of the test with a specified criterion when the total testing time is fixed. In the present work, the results obtained from an application of that method will be compared with the results obtained from an application of an *analytic method* proposed by Jackson and Novick (1970).

The General Context

The situation is one in which a test is composed of n subtests. The test score is the sum of the subtest scores, where a subtest score is computed as the number-right score.

Suppose T is the total time available for testing. We assume that we can shorten or lengthen each of the subtests. We wish to determine the amount of time t_i, $i=1,2,\ldots,n$, where $\sum t_i = T$ is fixed, to be allotted to each subtest (the number of items in each

subtest is determined by the time allotted to that subtest and the latency per item for that subtest) so that the predictive validity of the test score is maximized.

The Search Method

The process of finding the allocation of testing time which would maximize the predictive validity of the test score, under the constraint that $\sum t_i = T$, is based on calculating the predictive validity of the test score that would be obtained under different allocations of T and identifying the allocation that yields the highest validity. The predictive validity of the test score (X) with respect to a criterion (Y) for a given allocation of T among its n subtests is:

$$(1)\ r_{XY} = \frac{\sum r_{X_i(t_i)Y} s_{X_i(t_i)}}{\sqrt{\sum s_{X_i(t_i)}^{\ 2} + 2\sum r_{X_i(t_i)X_j(t_j)} s_{X_i(t_i)} s_{X_j(t_j)}}}$$ (Guilford, 1965, p. 427)

In equation (1) each subtest is allotted a certain amount of testing time, t_i, and $X_i(t_i)$ is the score of subtest i at that amount of time. The test score is thus defined as $X = \sum X_i(t_i)$.

$r_{X_i(t_i)Y}$ and $s_{X_i(t_i)}$ are the validity and the standard deviation respectively of subtest i allocated t_i testing time, and $r_{X_i(t_i)X_j(t_j)}$ is the intercorrelation between two subtests, i and j (when $j>i$), allocated a testing time of t_i and t_j respectively.

The values of the components on the right side of equation (1) can be computed directly for the *existing* ("initial") allocation of testing time in a given test, using the scores obtained by a group of examinees. Based on these values, as well as on the initial reliabilities (with $r_{X_iX_i}$ denoting the initial reliability of subtest i) of the subtests, the values that would be obtained under any *hypothetical* allocation of testing time can be calculated, using information regarding the typical latency per item for each subtest. Specifically, given such information, the ratio (K_i) between the number of items which are completed during any amount of time (t_i) allotted to the subtest and the initial number of items in the subtest can be computed. Then, the desired values can be computed as follows for any hypothetical vector of t_i's [1]:

[1] r_{X_iY}, s_{X_i}, $r_{X_iX_j}$, $r_{X_iX_i}$ and $r_{X_jX_j}$ are the **initial** values for the validity, standard deviation, intercorrelation and reliabilities respectively.

(2) $r_{X_i(t_i)Y} = \dfrac{r_{X_iY}\sqrt{K_i}}{\sqrt{1+(K_i-1)r_{X_iX_i}}}$ (Gulliksen, 1950, p. 89),

(3) $s_{X_i(t_i)} = s_{X_i}\sqrt{K_i + K_i(K_i-1)r_{X_iX_i}}$ (Gulliksen, 1950, p. 71) and

(4) $r_{X_i(t_i)X_j(t_j)} = \dfrac{r_{X_iX_j}}{\sqrt{1/K_i + (1-1/K_i)r_{X_iX_i}}\sqrt{1/K_j + (1-1/K_j)r_{X_jX_j}}}$

(Gulliksen, 1950, p. 98)

Following the estimation of r_{XY} for different allocations of testing time, a search for the allocation that yields the highest validity is conducted[2].

The Analytic Method

Jackson and Novick offer a formal solution to the problem. This solution is valid and unique whenever it specifies nonnegative times for all subtests.

The Formal Solution

As before, $X_i(t_i)$ is defined as the observed score of subtest i allocated time t_i. Then, under the classical test theory model,

(5) $X_i(t_i) = t_iT_i + E_i(t_i)$,

where $E_i(t_i)$ is the error score for time t_i and T_i is the true score at unit length. A formal solution for the vector of time allocations (**t**) and the coefficient for the regression of Y on X (β) which maximizes the correlation between X and the true score of the criterion Y, subject to $\sum t_i = T$, is sought for. The solution is obtained by minimizing the expected squared errors of prediction with respect to β and **t**, subject to the stated constraint.

The explicit solution for the vector of time allocations thus obtained[3] is

(6) $\mathbf{t}^* = \mathbf{H}\left(\dfrac{1}{\beta^*}\boldsymbol{\delta} - \dfrac{1}{2}\mathbf{Ae}\right) + \dfrac{\mathbf{G}^{-1}\mathbf{e}}{\mathbf{e}'\mathbf{G}^{-1}\mathbf{e}}T$,

[2] A detailed description of the search process is provided in the "Methods and Techniques" section.
[3] See Jackson and Novick (1970) for the proof.

where β^*, the coefficient for the regression of Y on X, is given by

$$(7)\ \beta^* = \frac{T\mathbf{e}'\mathbf{G}^{-1}\boldsymbol{\delta} - \frac{1}{2}\mathbf{e}'\mathbf{G}^{-1}\mathbf{e}\mathbf{e}'\mathbf{A}\mathbf{H}\boldsymbol{\delta}}{T^2 + T\mathbf{e}'\mathbf{A}\mathbf{G}^{-1}\mathbf{e} - \frac{1}{4}\mathbf{e}'\mathbf{G}^{-1}\mathbf{e}\mathbf{e}'\mathbf{A}\mathbf{H}\mathbf{A}\mathbf{e}}.$$

Equations (6) and (7) include the following elements:

$\mathbf{G}$ = the variance-covariance matrix of T_i,

$\boldsymbol{\delta}$ = the vector of covariances of T_i with the true score of the criterion Y,

$\mathbf{A}$ = a diagonal matrix whose iith term is $a_i^2 = \sigma^2\left[E_i(1)\right]$,

$\mathbf{e}$ = {1,1,…, 1} and

$$\mathbf{H} = \mathbf{G}^{-1} - \frac{\mathbf{G}^{-1}\mathbf{e}\mathbf{e}'\mathbf{G}^{-1}}{\mathbf{e}'\mathbf{G}^{-1}\mathbf{e}}.$$

It should be noted that since T_i is defined as a true score on a subtest 1 unit of time long, the matrices $\mathbf{G}$ and $\boldsymbol{\delta}$, as well as matrices derived from them, relate to a condition where each subtest is 1 unit long.

If one or more of the elements of $\mathbf{t}^*$ are negative, then the solution obtained from (6) is invalid. It is not possible to obtain the correct solution by assigning zero times to these subtests and making a proportional adjustment in the other subtest lengths. A stepdown procedure (a backward allocation) is suggested for these cases.

A Backward Allocation Procedure

In order to apply the backward allocation procedure, an assumption is needed that the partial regression coefficients of the criterion on the true scores of the subtests are all positive. Should this be false, Jackson and Novick suggest that the variables having negative coefficients be reflected, so that the assumption is satisfied. Jackson and Novick show that, in these circumstances, if T is sufficiently large, then each predictor will receive a positive time allocation.

Such an allocation forms the starting point for the backward allocation procedure. Now T is allowed to decrease until some element of $\mathbf{t}^*$ becomes zero. The

corresponding subtest is eliminated from the predictor set and a (valid) solution is obtained for the remaining (n-1) subtests.

These steps are repeated until T is decreased to the value stated in the constraint.

As Jackson and Novick remark, this procedure does not necessarily provide an optimal allocation.

Data

The methods presented above are implemented and compared using data from 4,321 first-year students studying in 355 academic departments in six Israeli universities during the academic year 1997/98. All the students took one of two parallel Psychometric Entrance Test (PET) Hebrew versions. They were selected on condition that at least three students in their academic department were tested on the same form of PET. The PET is used in admissions to institutions of higher education in Israel (see Beller, 1994). PET is designed to assess abilities in three domains: verbal reasoning, quantitative reasoning and proficiency in English as a foreign language.

The Verbal reasoning domain consists of six subtests[4]: Words and Expressions (W&E), Analogies (Ana), Sentence Completions (SC), Letter-Exchange (LE), Logic (Log) and Reading Comprehension (RC).

The Quantitative reasoning domain consists of three subtests: Questions and Problems (Q&P), Graph or Table Comprehension (G&T) and Quantitative Comparisons (QC).

The English domain consists of three subtests: Sentence Completions (SC-E), Restatements (Res) and Reading Comprehension (RC-E).

PET consists of six sections, two per domain. Each section contains 25-30 multiple-choice items that are to be answered within 25 minutes (the numbers of items per subtest across both sections of each domain are presented in Table 1).

The predictor variables were the twelve subtests of PET.

The criterion was grade-point average (GPA) at the end of the first year of university studies.

[4] A description of the subtests may be found in Kennet-Cohen, Bronner and Cohen (2003).

Methods and Techniques

The following statistics were computed[5] for the scores on each of the subtests for the initial time allocation:

(a) Variance

(b) Reliability (estimated by the split-half method)[6]

(c) Validity (computed as a correlation with GPA)

(d) Intercorrelations with the other subtests

All the above statistics were corrected for range restriction using the correction for univariate selection in the three-variable case (Gulliksen, 1950, pp. 145-156). A composite score, consisting of an equally weighted PET score and a measure of high-school achievement, served as the explicit selection variable. Its standard deviation in an unselected sample was estimated by a weighted average of its standard deviation among applicants to an academic department (by university and academic year). These estimates were based on data for applicants to all Israeli universities during the academic years 1991/92 and 1992/93.

Mean latencies per item in each subtest were estimated from an experimental computer-based version of PET (Cohen, Ben-Simon, Moshinsky, & Eitan, 2002).

The initial statistics and the mean latencies for the subtests are presented in Table 1.

[5] Each was a weighted average of the respective values calculated by academic departments, universities and PET versions.

[6] The original estimates of the reliabilities were increased by 7% in order to guarantee that the variance-covariance matrix of the true-score variables (matrix **G**) would have a nonnegative determinant. This correction had to be made so that the *analytic method* could be applied. Details regarding the rationale for and consequences of this procedure are presented in Appendix 2.

Table 1

Number of Items, Variances, Reliabilities, Validities, and Intercorrelations of Subtests for the Initial Allocation of Testing Time, and Mean Latencies (in Seconds)

		W&E	Ana	SC	LE	Log	RC	Q&P	G&T	QC	SC-E	Res	RC-E
No. of Items		8	12	10	8	12	10	30	8	12	22	12	20
Variance		2.73	4.77	4.06	2.80	5.97	5.27	18.14	4.13	5.21	17.2	5.17	15.3
Reliability		0.52	0.57	0.58	0.61	0.65	0.71	0.73	0.60	0.48	0.78	0.57	0.79
Validity		0.17	0.24	0.19	0.16	0.20	0.20	0.31	0.15	0.25	0.13	0.16	0.20
Intercorrelations	Ana	0.50											
	SC	0.37	0.47										
	LE	0.44	0.48	0.45									
	Log	0.29	0.39	0.41	0.37								
	RC	0.37	0.43	0.43	0.45	0.44							
	Q&P	0.27	0.39	0.33	0.30	0.48	0.34						
	G&T	0.22	0.28	0.26	0.25	0.40	0.35	0.40					
	QC	0.19	0.29	0.26	0.21	0.35	0.25	0.54	0.28				
	SC-E	0.34	0.36	0.33	0.33	0.27	0.33	0.26	0.22	0.18			
	Res	0.31	0.37	0.39	0.34	0.34	0.36	0.32	0.23	0.24	0.64		
	RC-E	0.30	0.37	0.38	0.35	0.36	0.43	0.33	0.32	0.25	0.64	0.62	
Mean Latencies		16.9	29.3	51.5	47.0	73.7	73.7	62.7	67.6	48.0	34.9	58.2	76.6

Application of the Search Method

1. Given statistics (a)-(d), as well as the mean latencies for the subtests, equations (1)-(4) were used to estimate the predictive validity of the test score under different allocations of testing time.

2. A search for the allocation that yielded the highest validity started with an even allocation of the total time among the *n* subtests.

From this starting point, the procedure proceeded in steps. In the first step the predictive validity of the test score in each of *n* possible conditions of **subtracting** 1 minute of testing time from one of the *n* subtests was calculated. The condition in which the decline in the predictive validity of the test score was the lowest was retained. Then, the predictive validity of the test score in each of *n* possible conditions of **adding** that 1 minute of testing time to one of the *n* subtests was calculated. The condition in which the gain in the predictive validity of the test score was the highest was retained. These steps were repeated until a stabilized allocation of testing time among the subtests was obtained, where the subtest which was chosen to lose 1 minute of testing time was the same as the one that was subsequently chosen to gain back that 1 minute. In other words, the steps just described continued to the point

where no gain in the predictive validity of the test score could be obtained by transferring 1 minute of testing time from one subtest to another.

Application of the Analytic Method

1. Given statistics (a)-(d), as well as the mean latencies for the subtests, the elements on the right side of equations (6) and (7) were computed, with T_i defined as the true score on a subtest 1 minute long.

2. Using equation (6), **t*** was computed.

3. If one or more elements of **t*** was negative, a backward allocation procedure was applied.

Both methods were applied to the whole test (T=150) and to each of the three domains (T=50) separately.

Results and Discussion

The allocation of the total testing time among the subtests and the predictive validity of the score obtained from this allocation are presented in Table 2 for the *search* and the *analytic* methods. Parallel data are also presented for the initial allocation of testing time.

Table 2

The Allocation of Testing Time Among Subtests and the Predictive Validity of the Resulting Score Obtained Initially, by the *Search Method* and by the *Analytic Method*

	Testing Time (in minutes)												
	W&E	Ana	SC	LE	Log	RC	Q&P	G&T	QC	SC-E	Res	RC-E	Validity
The Whole Test													
Initial	2	6	9	6	15	12	31	9	10	13	12	25	0.383
Search	0	39	0	0	0	0	47	0	46	0	0	18	0.444
Analytic	25*	110	0	15*	0	0	0	0	0	0	0	0	0.415
The Verbal Domain													
Initial	2	6	9	6	15	12							0.320
Search	0	40	0	0	10	0							0.370
Analytic	5*	45	0	0	0	0							0.372
The Quantitative Domain													
Initial							31	9	10				0.363
Search							29	0	21				0.374
Analytic							29	0	21				0.374
The English Domain													
Initial										13	12	25	0.248
Search										0	14	36	0.263
Analytic										0	14	36	0.263

* The subtest was reflected.

As can be discerned in Table 2, changing the allocation of the testing time among the subtests, regardless of the method adopted, can increase the validity of the test. In some cases (e.g., in the Quantitative domain) the effect is marginal (3%), but in others (e.g., in the Verbal domain) it is quite substantial (16%). Needless to say, the size of the effect depends on the degree of similarity between the initial allocation and the one obtained through a validity maximization process. In this sense, the size of the effect testifies to the quality, in terms of predictive validity, of the existing allocation.

Turning to a comparison between the two validity maximization methods, it should first be noted that in each of the four cases the application of the *analytic method* involved the use of the backward allocation procedure (as is evidenced by the fact that in each of the four cases some allocations are zero). As was mentioned before, the rationale for this procedure rests on the assumption that the partial regression coefficients of the criterion on the true scores of the subtests ($\mathbf{G}^{-1}\boldsymbol{\delta}$) are all positive.

In order to satisfy this condition it was necessary to reflect subtests for which the partial regression coefficient was negative (see the subtests marked by asterisks). This is clearly not a viable option in practice: we would hardly be prepared to subtract rather than add the score on a subtest or, equivalently, to score correct answers zero and incorrect answers one in computing the total score. However, for the purpose of the present study, this option was admissible, thus lending, in a sense, an additional "degree of freedom" to the *analytic method* compared to the *search method*. This advantage proves itself in the case of the Verbal domain, where the allocation obtained by the *analytic method* results in a slightly higher validity than the allocation obtained by the *search method*. Contrary to this, in the case of the whole test, the solution offered by the *analytic method* is clearly inferior to the one proposed by the *search method*. This complex case is probably an example of the reservation mentioned by Jackson and Novick regarding the fact that the backward allocation procedure cannot be guaranteed to produce the absolute maximum correlation attainable. The procedure is optimal for the subtests present at a given stage, but there may be a subgroup of subtests, different from the one to which the procedure led, which would result in a higher correlation. It is reasonable to assume that this limitation of the backward allocation procedure is more compelling in situations where many subtests relating to divergent content areas are involved. As was demonstrated here, such cases may occur in practical work.

The results of the two methods are identical with respect to both the Quantitative and the English domains. In addition, it should be noted that when the *analytic method* was applied to an initial subgroup of subtests consisting only of the subtests which were included in the solution obtained by the *search method*, the solution (i.e., the allocation of testing time) obtained was identical to the one obtained by the *search method*. In other words, in a context which did not require an application of the backward allocation procedure (i.e., when the formal solution yielded $\mathbf{t}^*$ whose elements were all nonnegative), the results of the two methods were identical. Such a result validates the *search method* proposed here, since the formal solution offered by Jackson and Novick, contrary to the backward allocation procedure, is an algorithm guaranteeing an absolute maximum.

There is one other content-oriented finding which emerged clearly in the current application and which we feel is worth mentioning. It is interesting to note that the optimal allocation of testing time among the subtests implies allocating a substantial portion of the total testing time to the subtest *Analogies*. This finding deserves some attention given the decision, implemented in the New SAT, to eliminate *Analogies* from the test. That decision was no doubt obtained in the context of a very complex combination of considerations, with predictive validity being just one in many. This seeming incompatibility between the current results and decisions implemeted in a practical situation of high-stakes testing testifies to the complexity of the process of test development.

Conclusions

The work in this paper is of substantial practical value to test constructors who wish to determine the optimum relative lengths for subtests of a test. The *search method* was both validated by the *analytic method* proposed by Jackson and Novick (when their formal solution was applied) and overcame some of the *analytic method*'s limitations (when the backward allocation procedure was applied). These limitations are likely to be encountered in practical contexts. In addition, inaccuracies in estimation (see Appendix 2) which resulted in violations of the classical test theory model led to the virtual collapse of the *analytic method*, while the *search method* was found robust with respect to such mishaps.

References

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practice, 13*(2), 12-20.

Cohen, Y., Ben-Simon, A., Moshinsky, A., & Eitan, M. (2002). *Computer based testing (CBT) in the service of test accommodations*. A paper presented at the 28th annual IAEA conference, Hong Kong.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.{Reprinted in 1987. Hillsdale, NJ: Erlbaum.}

Jackson, P. H., & Novick, M. R. (1970). Maximizing the validity of a unit-weight composite as a function of relative component lengths with a fixed total testing time. *Psychometrika*, *35*, 333-347.

Kennet-Cohen, T., Bronner S., & Cohen Y. (2003). *Improving the predictive validity of a test: A time-efficient perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Woodbury, M. A., & Novick, M. R. (1968). Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. *Journal of Mathematical Psychology*, *5*, 242-259.

Appendix 1:
Validity Maximization by Determining Subtest Lengths Compared with Validity Maximization by Using Regression Weights

A question often raised is whether the somewhat unconventional notion and nontrivial application of maximizing validity of a test by determining optimal subtest lengths (followed by combining item scores with unit weights) has any advantage over a rather common and readily available approach to maximizing validity - that of leaving the subtests at their given (initial) lengths and using *regression weights* in combining the subtest scores.

In what follows the results of a comparison between the two approaches will be presented. The approach which seeks to determine optimal subtest lengths will be represented by the *search method*.

As was presented before, the correlation of the test at the initial lengths, using unit weights for the items was 0.383 (see the first row of Table 2). Changing the subtest lengths via the *search method* and, again, using unit weights for the items, led to a correlation of 0.444 (see the second row of Table 2). Now, if *regression weights* (for the subtests) are used with the initial lengths, the multiple correlation obtained is 0.417. Such results do testify that there is a certain, not very substantial, advantage to searching for an optimal allocation of the testing time among the subtests (and using unit weights for the items) compared to using *regression weights* for the subtests at the initial allocation of the testing time. However, we claim that the non-substantiality of this advantage results from the fact that the initial allocation is in a sense appropriate. In other words, it seems that the test constructers had some intuitive knowledge regarding the appropriate allocation of the testing time. This initial appropriateness makes the *regression weights approach* look comparably effective. But what happens if the initial allocation of the testing time is a shot wide of the mark? We examined such a situation by simulating a hypothetical allocation of the total testing time consisting of 139 minutes of testing time allotted to the subtest *Log* and 1 minute of testing time allotted to each of the other subtests. The results obtained in this hypothetical situation ("poor allocation"), with respect to both the *search method* and the *regression weights approach*, are presented in Table 3. The predictive validity obtained with the initial allocation (using unit weights for the items) is also

presented. These three values can be compared with the results obtained with the real test ("real allocation") which were described above.

Table 3

The Predictive Validity of the Test
with Two Initial Allocations
Obtained Initially and Subsequent to Applying
the *Regression Weights Approach* or the *Search Method*

	Initial Allocation	
	Real	Poor
Initially	0.383	0.297
Regression Weights	0.417	0.329
Search	0.444	0.444

As is clear from Table 3, the maximal correlation attainable through the *regression weights approach* depends on the initial allocation of the testing time. If the initial allocation is such that when using unit weights for the items the correlation is low (0.297) than applying *regression weights* for the subtests also results in a low multiple correlation (0.329). Contrary to this, the maximal value for the predictive validity obtained by the *search method* does not depend on the suitability of the initial allocation. This is a significant advantage of the *search method*.

A closely related issue pertaining to the multiple regression approach relates to the potential contribution resulting from **combining** multiple regression with an optimal allocation of the total testing time (rather than **comparing** the multiple regression approach with one method or another for optimally allocating the total testing time, as was described hitherto).

In other words, the question is whether the predictive validity of the test can be improved by optimally allocating the total testing time, assuming that regression weights**,** rather than unit weights, are used to determine the composite predictor. In an earlier paper, Woodbury and Novick (1968) studied the first option. Later, Jackson and Novick (1970) compared this option ("regression weight allocation") with the case assumed throughout the present work ("unit weight allocation"). They state that the multiple correlation obtained in the former case is an upper bound on the

correlation obtained in the latter, and they reason this by the fact that the backward allocation is an algorithm in the former, but not in the latter. A question remains whether the superiority of the regression weight allocation holds also in a situation where the formal solution suffices, so that there is no need to turn to the backward allocation procedure. This question can probably be addressed analytically. However, it can also be addressed by the *search method*, since the *search method* does not rely on a backward allocation procedure when some of the subtests need to be omitted from the solution. This issue deserves further consideration.

Appendix 2:
The Rationale for Adjusting the Estimates of the Reliabilities

The Context

Our initial application of the *analytic method* yielded results – for example, the results obtained with respect to the whole test - which raised concerns that a problem existed in the estimation of some input data.

Specifically, the solution obtained by the *analytic method* consisted of two subtests, both from the English domain (24 minutes to *SC-E*, with its scores reflected, and 126 minutes to *Res*), yielding a validity coefficient of **0.27**. Contrary to this, the solution obtained through the *search method* consisted of four subtests, one from the Verbal domain (45 minutes to *Ana*), two from the Quantitative domain (27 minutes to *Q&P* and 63 minutes to *QC*) and one from the English domain (15 minutes to *RC-E*), yielding a validity coefficient of **0.45**. These results clearly show the disadvantage of the solution when obtained by the *analytic method* compared to the solution obtained by the *search method*. Such a profound disadvantage cannot be explained away by the suboptimality of the backward allocation procedure (bearing in mind that the **initial allocation** yielded a validity coefficient of **0.38**). Furthermore, when the two subtests which were selected through the *analytic method* (as aforesaid, 24 minutes to *SC-E*, with its scores reflected, and 126 minutes to *Res*) were offered as the starting point for the *search method*, the solution obtained (150 minutes to *Res*) yielded a validity coefficient (0.271) higher than the one obtained through the *analytic method* (0.270). Such a result testifies to the fact that the solution obtained by the *analytic method* in this case was invalid, since although the backward allocation procedure cannot guarantee optimality throughout (i.e., an absolute maximum), its solution should be optimal for the *i* subtests considered

The Problem

Following a careful examination of the elements included in equations (6) and (7), we noticed that the variance-covariance matrix of the true scores on the subtests (the matrix **G**) had, in the combination of all the subtests and in many partial combinations of the subtests, a (marginally) negative determinant. This is clearly not a viable

characteristic of a variance-covariance matrix. Such a result indicates some inaccuracy in the estimation of the matrix components.

The variance-covariance matrix of the true scores on the subtests (the matrix **G**) is computed as follows:

(8) $\mathbf{G} = \mathbf{D}_t^{-1}\left(\mathbf{\Sigma}\mathbf{D}_t^{-1} - \mathbf{D}^{-2}\right)$, where

$\mathbf{D}_t$ = a diagonal matrix of initial testing times,

$\mathbf{\Sigma}$ = the variance-covariance matrix of the observed scores on the subtests at the initial testing times, and

D = a diagonal matrix with elements a_i^{-1}, where $a_i = \sigma\left[E_i(1)\right]$, that is, the standard error of measurement for 1 minute of testing time.

The elements in matrices $\mathbf{D}_t$ and $\mathbf{\Sigma}$ are both obtained directly from the observed data and, therefore, would not seem to be the source of the negative determinant. However, the elements in the matrix **D** are calculated from estimates of the reliabilities of the subtests (for the initial time allocation). The reliability estimates were obtained, as mentioned in the text, by the split-half method. As is well known, such a procedure does not yield a unique estimate of the test's reliability coefficient. Specifically, a given division of the test into halves can yield an underestimation of the reliability coefficient, leading to a negative determinant of the variance-covariance matrix of the true score variables. There was no reason to assume *a priori* that our estimates of the reliabilities were biased (the method for dividing the subtests into halves was odd-even); however, such an outcome is a possibility. In addition, it should be noted that the restrictions we imposed on the number of observations (at least three) needed for the computations of within-department statistics led to a situation where not all the departments involved in computing the elements of the matrix $\mathbf{\Sigma}$ were involved in computing the reliabilities. Specifically, while 355 departments were involved in computing the elements of $\mathbf{\Sigma}$, only a subset of 274 departments were involved in computing the elements of **D**. Thus, minor inconsistencies between the estimations may have been obtained just because the sources of the estimations were not identical.

The Solution

It was decided to increase the reliabilities until a positive determinant of the variance-covariance matrix of the true score variables (the matrix **G**) was reached.

A 7% increase of the reliabilities guaranteed that this condition was maintained with respect to the variance-covariance matrices of the true scores of all the combinations of the subtests.

An Examination of the Potential Consequences of Correcting the Estimates of the Reliabilities

The consequences of correcting the estimates of the reliabilities were examined with respect to the *search method*. In other words, the results obtained by this method, based on the original estimates of the reliabilities, were compared with the results obtained after the correction of these estimates.

Table 4 presents the allocation of the total testing time among the subtests and the predictive validity of the score obtained by the *search method* for the original and the corrected estimates of the reliabilities.

Table 4

The Allocation of Testing Time Among Subtests and the Predictive Validity of the Resulting Score Obtained by the *Search Method* for the Original and the Corrected Estimates of the Subtest Reliabilities

Estimated Subtest Reliabilities	Testing Time (in minutes)												Validity
	W&E	Ana	SC	LE	Log	RC	Q&P	G&T	QC	SC-E	Res	RC-E	
Original	0	45	0	0	0	0	27	0	63	0	0	15	0.448
Corrected	0	39	0	0	0	0	47	0	46	0	0	18	0.444

The results presented in Table 4 testify to the fact that the correction of the estimates of the reliability had a negligible effect on the results obtained by the *search method*. Thus, the *search method* does not seem sensitive to the kind of inconsistencies exhibited in the data presented above.

In summary, two points are worthy of mention. First, with respect to both the *search* and the *analytic method*s, raising the estimates of the initial reliabilities lowers the potential gain in predictive validity obtained through any changes in the allocation of testing time. This can be clearly discerned in Table 4 (by comparing the two values appearing in the last column). Thus, this correction works to the disadvantage of both

methods proposed for increasing the validity. Second, in comparing the methods, it can be concluded that while the *analytic method* is sensitive to minor inaccuracies in estimation, the *search method* is more robust.