Improving the Predictive Validity of a Test: A Time-Efficient Perspective

Tamar Kennet-Cohen, Shmuel Bronner and Yoav Cohen

Paper presented at the annual meeting of the
National Council on Measurement in Education, Chicago, 2003

We would like to thank Avi Allalouf, Anat Ben-Simon, Ziva Canaan-Yehoshafat, Ruth Fortus, Naomi Gafni, Carmel Oren, Yosef Pavlov, Bella Pavzner-Serri, Joel Rapp and Reuven Stoller for their time, goodwill and indispensable contributions.

Abstract

Tests used for college or university admissions normally contain several types of items. After the desired set of item types has been specified, a decision regarding the proportions of the various item types has to be made. This work offers an approach for determining these proportions. The proposed approach is based on maximization of the predictive validity of the total score with respect to success in higher education, under the constraint of the total testing time available. A procedure of searching for the best allocation of the total testing time among the various item types is presented. This procedure makes use of statistical characteristics of the item types, such as reliability, validity, intercorrelations and variance, coupled with data regarding the response latencies for the item types. The proposed procedure can accommodate additional considerations regarding the desired proportions of the item types by introducing them in the form of additional constraints on the solution.
An application of the proposed approach is presented, based on data obtained for 4,543 first-year students in universities in Israel, to whom the Psychometric Entrance Test was administered.

Introduction

A major step in the process of test construction, once the behaviors that represent the construct to be measured have been identified, is preparing test specifications. This process consists of determining a number of classification dimensions, among them the item types to appear on the test. Once the item types have been chosen, the issue becomes: how many items on the test should represent each item type?

In the context of a test designed to measure the outcome of a specific program of instruction, the number of items allotted to each item type usually reflects the amount of time devoted to mastery of material tapped by this item type during the course of instruction (Crocker & Algina, 1986). Contrary to this, in the context of a test of aptitude, which is generally curriculum free, there is usually no well-defined body of content to which to refer. Consequently, it is less clear how many items must be written for each type in order to represent the body of content adequately (Thorndike, 1982). Given these circumstances, it is proposed here that instead of deriving the specific frequencies for the item types using subjective judgement only, an explicit rationale be adopted. Such a rationale is exemplified in this work in the context of an aptitude test intended to predict future academic performance.

The proposed approach is basically based on the idea that the decision regarding the number of items allotted to each type should be geared to the purpose to be served by the test. Therefore, in the present context, we propose to view the decision faced by the test developer regarding the number of items allotted to each type as a problem of maximization of criterion-related validity under certain constraints. In other words, the number of items allotted to each type should be such that the predictive validity of the test with respect to the criterion of academic performance will be maximized, given certain explicit and quantifiable constraints.

The most obvious constraint on the test is that of time. It is well known that the length of a test has an effect on the test's reliability and validity. There is a

positive correlation between the number of items on a test and its reliability and validity. However, time considerations constrain the number of items that can be used on a test. Large-scale assessment usually requires that the test be administered during the course of one day. The factor of fatigue must also be taken into account: requiring examinees to reason or write for extended periods of time will result in a decline in performance. Thus, given that the time allotted for the test is limited, the considerations regarding the distribution of items among item types should take into account not only the contribution of an additional item of a given type to the validity of the test, but also its demands on the resource of total testing time, i.e., the response time it requires. This conceptualization leads to focusing on the validity (and on the contribution to the validity of the total test) of different item types **per unit of testing time,** and not per item.

The application of the approach described above involves, in general terms, the following steps: first of all, for a given number of items for each type in the test under consideration, the current reliability, predictive validity and some additional statistics for each item type are estimated. Secondly, data regarding response latency for each item type are obtained. Thirdly, the above information is used to estimate the statistics of interest for each item type if different amounts of testing time (1 minute, 2 minutes, etc.) are devoted to the item type. Finally, based on the information obtained in the previous step, it is possible to obtain the best (in terms of maximum predictive validity of the total score) allocation of the total testing time among item types.

This approach can be generalized to include additional constraints on the test design. Thus, for example, minimal and/or maximal limits on the number of items of a given type can be imposed because of psychometric as well as other considerations. These may include problems in constructing items of a certain type (at a desired level of difficulty), trying to avoid monotony and boredom, considerations of face validity, etc. The proposed approach is consistent with other approaches presented in the optimal test assembly literature (van der Linden, 1998a), which views test assembly as an optimization problem. Most notably, all these approaches share the basic

structure of constrained combinatorial optimization, wherein the task is to select a combination of items that is optimal with respect to one attribute and that meets a variety of constraints on other attributes (van der Linden, 1998b). The approach proposed here expresses the familiar notions of predictive validity and testing time in terms of an objective function and a set of constraints, thus demonstrating an innovative application of an optimization technique to the test assembly problem.

The proposed approach will be exemplified here with respect to a university entrance exam in Israel. University applicants in Israel were required, until recently, to submit scores on the Psychometric Entrance Test (PET) to universities. PET is designed to assess abilities in three domains: verbal reasoning, quantitative reasoning and proficiency in English as a foreign language. Our goal was to propose a method for deciding upon the allocation of the total testing time (150 minutes) among the various item types which currently appear on PET. This was carried out with the aim of maximizing the predictive validity of the number right score on the test with respect to the criterion of success (grade-point average) at the end of the first year of university studies.

Method

Sample

The analyses were based on data for 4,543 first-year students studying in 355 academic departments in six Israeli universities for the academic year 1997/98. All the students were tested in PET in one of two specific forms. They were selected on condition that at least three students in their academic department were tested on the same form of PET.

Variables

Predictors

Twelve item types are included in PET, which is constructed and administered by the National Institute for Testing and Evaluation (NITE). PET is designed to

assess abilities in three domains: verbal reasoning, quantitative reasoning and proficiency in English as a foreign language. The operational PET consists of six sections, two per domain. Each section contains 25-30 multiple-choice items of several types that are to be answered within 25 minutes (the numbers of items per type are presented in Table 1).

These are the types of items which appear in PET (Attali & Goldschmidt, 1999):

The domain of verbal reasoning (V)

1. *Words and Expressions (V-W&E)* : The purpose of this item type is to assess the vocabulary of examinees. The items appear in a number of forms: items dealing directly with the meaning of words and expressions; sentence-completion items with one blank; antonyms; and items in which examinees have to choose one word whose meaning is distinct from that of the other words.

2. *Analogies (V-Ana)*: Verbal analogies test the ability to define the relationship between two concepts, based on their semantic meanings, and to recognize a similar relationship in other pairs of concepts.

3. *Sentence-Completions (V-SC)*: Sentence-completion items consist of a sentence with three or four blanks. These items emphasize understanding of the logical and semantic relationships within a complex sentence.

4. *Letter-Exchange Items (V-LE)*: These items were developed at NITE and are based upon a morphological feature of Semitic languages not shared by Indo-European ones, namely, the fact that most of the vocabulary in Hebrew – all verbs and most nouns and adjectives – can be characterized as a combination of Root + Pattern. The root is most typically composed of three consonants, which denote the semantic core of the words formed by it; the patterns take the form of vocalic and syllabic additions to the root, which serve to modify the core meaning of the root.

The letter exchange items are composed of four sentences. In each sentence one word is altered by changing its root letters into a standard template (the letters **p.t.l.**). In three of the four sentences the standard template stands for the same three letters. In the remaining sentence the template replaces another root. The examinee has to identify this sentence.

5. *Logic (V-Lo)*: This item type includes deductive and inductive problems. There are two types of deductive problems. In one type, called categorical syllogisms, examinees are required to determine what conclusion, if any, must follow from certain assumptions about category membership. The other type, called propositional problems, concerns the evaluation of the truth of arguments consisting of sequences of simple statements linked by connectives such as *and*, *or*, *not* and *if…then* to form compound statements. The type of inductive problem included in *V-Lo* is hypothesis testing: examinees are required to perform a variety of tasks, such as judging the plausibility of conclusions, recognizing assumptions with respect to certain conclusions, or analyzing the effects of additional information on a conclusion.

6. *Reading Comprehension (V-RC)*: This item type consists of a short academic text followed by a number of questions. The items reflect the process that a reader goes through while deriving meaning from a text. They assess the test taker's ability to interpret, synthesize, analyze, and evaluate the reading material, and thus measure higher order analytical and evaluative skills.

The content of the reading selections is drawn from a variety of academic domains, for example, the humanities, biological and physical sciences, and social sciences.

The domain of quantitative reasoning (Q)

1. *Questions and Problems (Q-Q&P)*: These questions are divided into verbal questions, non-verbal questions and geometry. The verbal questions involve translating the problem into algebraic expressions. The non-verbal questions already contain algebraic expressions. The topics covered by verbal and non-verbal questions include equation solving, distance and work problems, combinatorial analysis and probability. The geometry questions deal with characteristics of geometrical shapes, such as area, volume, angles and the like.

2. *Graph Comprehension or Table Comprehension (Q-G&T)*: The information in these questions is presented in the form of a graph or a table. The table can be of one or more dimensions. The graph presents data in graphic form, such as a curve, a bar chart or a scatterplot. The questions are of two main types:

- Questions involving the reading of data, in which examinees are asked to find information appearing in the graph or table.
- Inference questions, in which examinees are asked to make various inferences based on the data appearing in the graph or table.

3. *Quantitative Comparisons (Q-QC)*: These items consist of pairs of quantities, with additional information sometimes provided. Based on the quantities and the additional information (if provided), examinees are asked to decide whether one of the quantities is larger than the other, whether the two quantities are equal, or whether not enough information has been provided to determine the relationship between the two quantities. The content covered by these items is the same as in *Q-Q&P*.

The domain of English (E)

1. *Sentence Completions (E-SC)*: Sentence Completion items consist of sentences with a word or words missing in each. Examinees are asked to choose the answer which best completes the sentence. These items test English vocabulary and the ability to use English words in a given context.

2. *Restatements (E-Res)*: In Restatement items, a sentence is presented, followed by four possible restatements of that sentence. For each question, examinees are asked to choose the one restatement which best expresses the meaning of the original sentence. These items are intended to test vocabulary, syntax, and the ability to understand the relationships between different parts of the sentence.

3. *Reading Comprehension (E-RC)*: Reading Comprehension items assess the ability to understand short academic texts. The items related to a text could touch upon a word, a sentence, or a larger part of the text.

The number-right score on each of the item types was computed across both sections of the domain. The total of scores on all the item types (*Tot*) was computed across all six sections of the test.

Criterion

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

Procedures

Stage 1

The following statistics were computed for the number-right scores on each of the item types:

- Variance
- Reliability: The reliability of each item type was estimated by the split-half (odd-even) method.
- Validity: The validity coefficient of each item type was computed by correlating it with GPA.

In addition, intercorrelations between item types were computed.

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). This adjustment requires using the standard deviation of the explicit selection variable in an unselected sample. PET total score was treated 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 school year). These estimates were based on data for applicants to all Israeli universities during the school years 1991/92 and 1992/93 (a detailed description of the correction method can be found in Kennet-Cohen, Bronner, & Oren, 1999). For the reliability coefficients, the correlation between the two halves of each item type was corrected for range restriction prior to using the Spearman-Brown formula.

All the above statistics were computed within academic departments (within university and PET form), weighted by the number of students in that department, and averaged across departments.

The final statistics obtained for Stage 1 are presented in Table 1.

Table 1

Number of Items, Variances, Reliabilities, Validities, and Intercorrelations of Item Types Used under Current Testing Conditions

		V-W&E	V-Ana	V-SC	V-LE	V-Log	V-RC	Q-Q&P	Q-G&T	Q-QC	E-SC	E-Res	E-RC
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		0.49	0.53	0.54	0.57	0.61	0.66	0.68	0.56	0.45	0.73	0.53	0.74
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
Inter-Corr.	V-Ana	0.50											
	V-SC	0.37	0.47										
	V-LE	0.44	0.48	0.45									
	V-Log	0.29	0.39	0.41	0.37								
	V-RC	0.37	0.43	0.43	0.45	0.44							
	Q-Q&P	0.27	0.39	0.33	0.30	0.48	0.34						
	Q-G&T	0.22	0.28	0.26	0.25	0.40	0.35	0.40					
	Q-QC	0.19	0.29	0.26	0.21	0.35	0.25	0.54	0.28				
	E-SC	0.34	0.36	0.33	0.33	0.27	0.33	0.26	0.22	0.18			
	E-Res	0.31	0.37	0.39	0.34	0.34	0.36	0.32	0.23	0.24	0.64		
	E-RC	0.30	0.37	0.38	0.35	0.36	0.43	0.33	0.32	0.25	0.64	0.62	

Stage 2

Information regarding the time needed to complete items of different types was estimated using data on average response times obtained from an experimental administration of a computerized linear (i.e., nonadaptive) administration of PET with a 30 minutes time constraint per section (Cohen, Ben-Simon, Moshinsky, & Eitan, 2002). However, some adjustment of the latencies was still necessary, because of two reasons. The main reason is that the operational paper and pencil (P&P) test is administered with a time limit of 25 minutes per section, whereas in the experimental computerized administration an extra 5 minutes were added to each section. The second reason is that in the experimental administration not all the items were reached and answered; the estimates mentioned above were based only on items which were answered. However, in the operational administrations of PET the vast majority of the examinees fill in all the answers on the answer sheet. Because of these two differences between the operational and the experimental administrations, the response latencies obtained from the experimental administration were adjusted in such a way that the total time needed to answer all the items in the domain would be 50 minutes. In order to

achieve this, the values of the latencies for each item type in each domain were multiplied by the ratio between the time limit for the domain in the operative administration (50 minutes for each domain) and the total time needed to answer all the questions in the domain, given the current number of items and the latencies obtained from the experimental administration (56 minutes, 61 minutes and 60 minutes for V, Q and E respectively).
One of the item types (*V-LE*) was not included in the above experiment. Therefore, in order to estimate its average response time, we had to use data from an experimental administration of a computerized adaptive version of PET (Moshinsky, 2000). These data included average response times for *V-LE* as well as for all the other item types. The response time for *V-LE*, adjusted for computerized linear administration, was extrapolated from the response time for *V-LE* relative to the response times for the other item types in the computerized adaptive administration.

The latencies obtained from the experimental computer based testing (CBT) and the latencies adjusted for the operational P&P administration of PET are presented in Table 2.

Table 2

Mean Latency (in Seconds) per Item in each Item Type

	V-W&E	V-Ana	V-SC	V-LE	V-Log	V-RC	Q-Q&P	Q-G&T	Q-QC	E-SC	E-Res	E-RC
CBT	19.0	33.0	58.0	53.0	83.0	83.0	77.0	83.0	59.0	42.0	70.0	92.0
P&P	16.9	29.3	51.5	47.0	73.7	73.7	62.7	67.6	48.0	34.9	58.2	76.6

Stage 3

Given the (adjusted) latencies for each item type which are presented in Table 2, and the statistics which are presented in Table 1, the variance, reliability and validity of each item type, as well as the intercorrelations between item types, can be computed for any allocation of testing time among the different item types.

Such computations will be illustrated here for the general case where t_i (i refers here to the item type, i varies from 1 to 12) minutes of testing time is devoted to each item type.

For each item type:
The number of items which can be completed in t_i minutes is computed as:

Number of items completed in t_i minutes = $\frac{t_i \times 60}{\text{latency}_i}$

The ratio of the number of items of type i which can be completed in t_i minutes of testing to the current number of items of type i is:

$K_i = \frac{\text{number of items in } t_i \text{ minutes}}{\text{current number of items}}$

The **variance** of the item type in t_i minutes of testing is:

$s^2_{K_i} = s^2_{C_i} K_i \left[1 + (K_i - 1) r_{CC_i}\right]$ (Gulliksen, 1950, p. 71)

where $s^2_{C_i}$ is the variance of the item type given the current number of items and

r_{CC_i} is the reliability of the item type given the current number of items.

The **reliability** of the item type in t_i minutes of testing is:

$r_{KK_i} = \frac{K_i r_{CC_i}}{1 + (K_i - 1) r_{CC_i}}$ (general Spearman-Brown formula, Gulliksen, 1950, p. 78)

The **validity** of the item type in t_i minutes of testing is:

$r_{K_i GPA} = \frac{r_{C_i GPA} \sqrt{r_{KK_i}}}{\sqrt{r_{CC_i}}}$ (Gulliksen, 1950, p. 89)

where $r_{C_i GPA}$ is the validity of the item type given the current number of items.

The **intercorrelation** between two item types, i and j, tested for t_i and t_j minutes respectively is:

$$r_{K_iK_j} = \frac{r_{C_iC_j}\sqrt{r_{KK\,i}}\sqrt{r_{KK\,j}}}{\sqrt{r_{CC\,i}}\sqrt{r_{CC\,j}}}$$ (Gulliksen, 1950, p. 98)

where $r_{C_iC_j}$ is the correlation between item types i and j, given the current number of items.

As an example, the statistics obtained for a situation in which 1 minute of testing time is devoted to each item type are presented in Table 3.

Table 3

Number of Items, Variances, Reliabilities, Validities, and Intercorrelations of Item Types with 1 Minute of Testing Time for each Item Type

		V-W&E	V-Ana	V-SC	V-LE	V-Log	V-RC	Q-Q&P	Q-G&T	Q-QC	E-SC	E-Res	E-RC
No. of Items		3.56	2.05	1.16	1.27	0.81	0.81	0.96	0.89	1.25	1.72	1.03	0.78
Variance		0.88	0.46	0.25	0.23	0.18	0.17	0.20	0.23	0.32	0.44	0.23	0.17
Reliability		0.30	0.16	0.12	0.17	0.09	0.14	0.06	0.12	0.08	0.17	0.09	0.10
Validity		0.14	0.13	0.09	0.09	0.08	0.09	0.10	0.07	0.10	0.06	0.06	0.07
Inter-Corr.	V-Ana	0.21											
	V-SC	0.14	0.12										
	V-LE	0.19	0.15	0.12									
	V-Log	0.09	0.08	0.08	0.08								
	V-RC	0.13	0.11	0.09	0.11	0.08							
	Q-Q&P	0.06	0.06	0.05	0.05	0.06	0.05						
	Q-G&T	0.08	0.07	0.06	0.06	0.07	0.08	0.06					
	Q-QC	0.06	0.07	0.05	0.05	0.06	0.05	0.07	0.06				
	E-SC	0.13	0.10	0.08	0.09	0.05	0.07	0.04	0.05	0.04			
	E-Res	0.10	0.08	0.08	0.08	0.05	0.07	0.04	0.04	0.04	0.13		
	E-RC	0.09	0.08	0.07	0.07	0.05	0.07	0.04	0.06	0.04	0.11	0.09	

Stage 4

At this stage, a search was conducted for the allocation of the total testing time (150 minutes) among the currently existing item types which would maximize the predictive validity of the composite score, computed as the total score on all the item types (*Tot*).

Thus, the expression to be maximized was:

$$r_{TotGPA} = \frac{\sum r_{K_iGPA}\, s_{K_i}}{\sqrt{\sum s_{K_i}{}^2 + 2\sum r_{K_iK_j}\, s_{K_i}\, s_{K_j}}}$$ (Guilford, 1965, p. 427)

when maximization was conducted under the constraint that

$\sum t_i = 150$

where, as before:

t_i is the number of minutes of testing time devoted to each item type

r_{K_iGPA} is the validity of an item type in t_i minutes of testing

s_{K_i} is the standard deviation of an item type in t_i minutes of testing

$r_{K_iK_j}$ is the intercorrelation between two item types, i and j (when j>i), tested for t_i and t_j minutes respectively.

The process of finding the allocation of testing time which would maximize the predictive validity of the composite score, r_{TotGPA}, under the constraint of a total testing time of 150 minutes, was based on defining different allocations of the testing time and examining the predictive validity of the composite score that would be obtained under each allocation. The end product of this process was the identification of the allocation that yielded the highest validity of the composite score.

The above process was conducted as follows:

The starting point was an even (or close to even) allocation of the total of 150 minutes of testing time among the item types.

From this starting point, the procedure proceeded in steps. In the first step the predictive validity of the composite score in each of 12 possible conditions of **subtracting** 1 minute of testing time from one of the 12 item types was calculated. The condition in which the decline in the predictive validity of the composite score was the lowest was retained. Then, the predictive validity of the composite score in each of 12 possible conditions of **adding** that 1 minute of testing time to one of the 12 item types, was calculated. The condition in which the gain in the predictive validity of the composite score was the highest was retained. These steps were repeated until a stabilized allocation of testing time among the item types was obtained, where the item type 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. Thus, the steps just described continued to the point where no gain in predictive validity of the

composite score could be obtained from transferring 1 minute of testing time from one item type to another.

Additional Constraints

In addition to the maximization problem to be solved under the constraint of a total of 150 minutes of testing time, four other maximization tasks were examined. These all had the same target: finding the allocation of testing time which maximized the validity of the composite score while varying the degree of restriction imposed on the solution.

These restrictions are described in Table 4, together with the restriction described above, i.e., the constraint of a total of 150 minutes of testing time (Condition 5 in the Table).

Table 4

Restrictions (in Minutes) Imposed on the Solution in the Validity Maximization Problem

Condition	Total Testing Time	Time per Domain	Upper Limit to Time per Item Type	
			Verbal Domain	Quantitative or English Domain
1	150	50	10	20
2	150	50	15	25
3	150	50	20	30
4	150	50	-	-
5	150	-	-	-

It can be seen that the degree of restriction decreases from Condition 1 to Condition 5, each condition being a more general case than the condition which precedes it.

In conditions 1 to 4, a restriction is imposed so that there will be a total of 50 minutes testing time per domain. This restriction corresponds to the time allotted to each domain in the operational administration of PET.

Conditions 1 to 3 impose, in addition to time per domain, varying degrees of constraints on the number of item types included within a domain: the lower the upper limit of testing time per item type, the larger the number of item types which will be included in the solution. Thus, for example, Condition 3 demands that at least three item types be included in the verbal domain,

Condition 2 demands that at least four item types be included in the verbal domain, and Condition 1 demands that at least five item types be included in the verbal domain. The same principle applies to the quantitative and English domains. The specific numbers which indicate the upper limits to time per item type were chosen in accord with the current number of item types in each domain.

It can generally be expected that the higher the degree of constraint imposed on the solution for the "best" allocation, the lower the predictive validity of the composite score which is derived from this allocation. Thus, the validity of the obtained composite score is expected to be the highest in Condition 5 and the lowest in Condition 1. The important question is, however, what price would have to be paid, in terms of predictive validity, for introducing any additional *a priori* demands on the composition of the test.

The predictive validity of the composite scores obtained in the five conditions can also be compared with the predictive validity of the current composite score. The predictive validity of the current composite score was obtained by using the formula for r_{TotGPA}, where the elements of the formula were computed with $k_i=1$ for all the item types.

Cross-Validation

The procedure described for deciding upon the allocation of the testing time among the item types is basically a method of optimization. Hence, a cross-validation study is called for in order to check the applicability of the obtained allocation of testing time in a new sample from the same population.
Thus, in addition to the analyses performed on the total sample, a holdout cross-validation procedure (Ghiselli, Campbell & Zedeck, 1981) was used. The current sample was randomly divided into two groups, controlling for some essential characteristics of the units of analysis: university, academic department and PET form. The procedures described in Stage 4 were applied to one of the groups - the experimental group (which included 2,174 students in 177 departments). The solution obtained for the allocation of the total

testing time among the item types (for each of the 5 constraining conditions) were applied in constructing composite scores for the second group – the holdout group (which included 2,369 students in 178 departments). The predictive validity of the composite scores (r_{TotGPA}) was computed in the holdout group in order to cross-validate the proposed allocation of the total testing time.

Results

The allocation of the total testing time among the twelve item types under each of the five conditions and the predictive validity of the composite score obtained from this allocation are presented in Table 5. Parallel data are also presented for the current allocation of testing time. The current allocation of testing time among the item types is calculated from the current number of items in each type.

Table 5

The Allocation of Testing Time (in Minutes) among Item Types and the Predictive Validity of the Resulting Composite Score under Current (C) and Five Maximization-under-Constraint Conditions (1-5)

	Testing Time (in minutes)												Validity
	V-W&E	V-Ana	V-SC	V-LE	V-Log	V-RC	Q-Q&P	Q-G&T	Q-QC	E-SC	E-Res	E-RC	(r_{TotGPA})
In the Total Sample													
C	2	6	9	6	15	12	31	9	10	13	12	25	0.31
1	6	10	10	4	10	10	20	10	20	10	20	20	0.32
2	-	15	5	-	15	15	25	-	25	-	25	25	0.34
3	-	20	-	-	20	10	27	-	23	-	20	30	0.34
4	-	47	-	-	3	-	27	-	23	-	-	50	0.36
5	-	41	-	-	-	-	20	-	89	-	-	-	0.38
Cross-Validation													
C	2	6	9	6	15	12	31	9	10	13	12	25	0.29
1	-	10	10	10	10	10	20	10	20	10	20	20	0.29
2	-	15	5	-	15	15	25	-	25	-	25	25	0.32
3	-	20	-	-	20	10	30	-	20	-	30	20	0.32
4	-	45	-	-	5	-	50	-	-	-	50	-	0.32
5	-	28	-	-	-	-	122	-	-	-	-	-	0.36

Note: The validities of the composite score for the experimental group in the cross-validation procedure were: 0.34, 0.34, 0.36, 0.36, 0.36 and 0.41 under the current and the five conditions respectively.

Data obtained for the total sample are presented in the upper panel of Table 5. Data obtained from the cross-validation procedure are presented in the lower panel.

The results, first with respect to those obtained in the total sample, indicate that a considerable improvement in the validity of *Tot* can be obtained by changing the allocation of the total testing time among the item types. Even when severe constraints are imposed on the composition of the test, such as in Condition 2, a gain of 10% in the predictive validity of *Tot* is observed. And, if some of these constraints can be removed, the gain in predictive validity is considerably larger.

Apart from the gain in predictive validity, a tentative picture emerges when examining the proposed allocation of the testing time in the various conditions. In the verbal domain there is a clear advantage for one item type - *V-Ana*. The less advantageous item types, in terms of predictive validity, are *V-W&E* and *V-LE*, which have the smallest contribution to the validity of the composite score. They are assigned testing time only under the most severe constraint (Condition 1).

In the quantitative domain the leading role is generally shared by two item types: *Q-Q&P* and *Q-QC*. The item type *Q-G&T* contributes the least to predictive validity, and is left out of the allocation as soon as the constraints permit.

In the English domain there is a clear advantage to *E-RC* and a clear disadvantage to *E-SC*.

An additional interesting finding is observed in Condition 5, where no constraints are imposed besides the one for a total testing time of 150 minutes. The proposed allocation of the testing time in this condition includes only items from the verbal and quantitative domains, and none from the English domain.

An examination of the results obtained in the cross-validation, and a comparison with the results obtained in the total sample, lead to the following impressions: A clear improvement in predictive validity can be discerned when a re-allocation of the testing time is implemented via a cross-validation procedure. Again, even in Condition 2, a gain of 10% in the predictive validity of *Tot* is obtained. In fact, the specific allocation of testing time obtained in

Condition 2 for the cross-validation is identical to the one obtained in the total sample. When the constraints are gradually removed, only minor inconsistencies between the results obtained in the cross-validation and those obtained in the total sample are found: in the verbal domain the results are practically identical, in the quantitative domain a clear advantage for item type *Q-Q&P* is obtained in the cross-validation (compared to a certain advantage for *Q-QC* in the total sample), and in the English domain a clear advantage for item type *E-Res* is obtained in the cross-validation (compared to an advantage for *E-RC* in the total sample). A definite similarity between the results obtained in the total sample and those obtained in the cross-validation can be observed with respect to those item types which were found to contribute the least to the predictive validity of the composite score (*V-W&E* and *V-LE* in the verbal domain, *Q-G&T* in the quantitative domain and *E-SC* in the English domain). Finally, as before, when no constraints are imposed besides the one for a total testing time of 150 minutes (Condition 5), the proposed allocation of the testing time includes only items from the verbal and quantitative domains, and none from the English domain. All in all, the results obtained in the cross-validation seem to testify to the generalizability of the results reported for the total sample.

Discussion

In summarizing and discussing the results presented above it should be kept in mind that the main purpose of this work was to propose a certain approach – an approach for deciding upon the allocation of testing time among item types such that the predictive validity of the test be maximized – and to demonstrate its implementation. As such, the potential contribution of this work should be evaluated more on the basis of the soundness, feasibility, effectiveness and practical utility of the method presented here, and less on the specific results obtained. However, given the rather clear and stable pattern of results, some tentative conclusions regarding the actual subject matter will be presented later.

The basis for the approach which was examined here was that predictive validity considerations should receive explicit and formal treatment when developing a test to be used for selection. Thus, the essence of the method proposed here is that in the process of test development, certain statistical characteristics of the item types – such as the correlation of each item type with the criterion and the covariances and variances of the item types – be combined with data regarding the response latencies for the various item types, in order to find the allocation of testing time which maximizes the predictive validity of the test. Combining the two types of information, statistical characteristics and response latencies, enables us – under the assumption that all the items which are added to or subtracted from a type are parallel – to compute the expected values of the relevant characteristics of the item types for different allocations of testing time. These expected values are then used in computing the predictive validity of the composite score according to the different allocations of total testing time. In the application which was presented here, a search process was conducted. This process started from a given allocation of the testing time. From this starting point, possible transitions of a unit of testing time from one item type to another were examined. The allocation of time that yielded the largest improvement in the predictive validity of the composite score was selected.

Although the study presented here should generally be viewed as a practical illustration of the proposed approach for deciding upon item type proportions in a given test, the results obtained seem worthy of discussion. The approach proved effective in the context of a real test: the predictive validity of a composite score based on the sum of the scores on all the item types was raised by more than 20% when no constraints (except for a total of 150 minutes of testing time) were imposed on the solution, and by 10% when rather severe constraints (that 50 minutes of testing time be devoted to each domain and that at least two thirds of the item types currently included in the test remain in it) were imposed.

As for the proposed allocation of testing time, the consistencies, as well as the minor inconsistencies, between the results obtained in the total sample and those obtained through cross-validation suggest the following: In the verbal domain there is a clear advantage to *analogies* (*V-Ana*); its share in the 50 minutes of testing time assigned to the verbal domain rose from 10% to more than 90% (in Condition 4). This item type (together with series problems and classifications) has played a key role in both the psychometric and the information-processing literatures on inductive reasoning (e.g., Sternberg, 1977, 1984). Sternberg, for example, tried to identify, through an analysis of people's reaction times and error rates in analogies and similar tasks, the "components of intelligence" – those mental processes and strategies used in information processing in complex tasks. In general, analogical reasoning is conceived of as a powerful tool in learning and understanding. It can be interesting to examine the results reported here regarding the contribution of different item types to the criterion-related validity of the test in light of the research (Gentile, Kessler & Gentile, 1969; Whitely & Barnes, 1979; Sternberg, 1982; Bejar, Chaffin & Embretson, 1991) dealing with the identification of the cognitive processes involved in the solution of analogies. The question that is often posed in this context is: Are analogies a verbal reasoning test, in the sense that their difficulty lies in the relations between the components, or are they a vocabulary test, in the sense that their difficulty lies in the lack of familiarity with the words used in the item? Past research (Roccas & Moshinsky, in press) has indicated that the difficulty of analogies is affected by both the complexity of the cognitive processes needed to define the relations between the words (i.e., reasoning) and the difficulty of the words (i.e., vocabulary). The fact that both these factors are related to analogies is indicated by the pattern of the intercorrelations among the item types, where the correlations of analogies with the vocabulary-oriented item types (*Words and Expressions – V-W&E* and *Letter-Exchange – V-LE*) are somewhat higher than the correlations with the reasoning-oriented item types (*Sentence-Completion – V_SC* and *Logic – V_Lo*). However, it seems that the factor related to analogies which is more relevant to the criterion is the reasoning component and not the vocabulary one. This can be clearly discerned by the fact that the vocabulary-oriented item types are allotted no testing time in the

solutions obtained, despite their clear advantage (in particular with regard to *V-W&E*) in terms of response latency. In contrast to this, the reasoning-oriented item types do seem to contribute to the predictive validity of the composite score, as can be discerned by the testing time allotted to them in the solutions obtained. However, given the rather short response latency for analogies, this item type gains a clear advantage over the two reasoning oriented item types.

In the quantitative domain an impressive stability is observed between the results obtained in the total sample and those obtained in cross-validation with regard to the item type which contributes the least – *Graph and Table Comprehension* (*Q-G&T*). This item type is allocated testing time only under the most severe constraint, which demands that all the item types which are currently included in the quantitative domain remain in it. The two other item types – *Questions and Problems* (*Q-Q&P*) and *Quantitative Comparisons* (*Q-QC*) – have a similar contribution to the validity of the composite score.

In the English domain, the item type which contributed the least was *Sentence Completions* (*E-SC*). This item type (unlike *Sentence Completions* in the verbal domain – *V-SC*) mainly measures vocabulary. Thus, like the results obtained in the verbal domain, it was found that item types which mainly measure vocabulary – be it in mother-tongue or in a foreign language – have a relatively low potential validity and no incremental, unique contribution to the validity of the composite score, despite their considerable advantage in terms of response latency. The two other item types in the English domain assess the ability to understand complex English sentences (*Restatements – E-Re*) or short texts (*Reading Comprehension - E-RC*). Their contributions to the validity of the composite score are similar to each other. In such circumstances (both in English and in the quantitative domain) it might be advisable to include both item types in the test, in order to obtain the gains achieved by diversity.

A final observation pertains to the results obtained in the least-constrained condition. When no restrictions were imposed on the solution, except for a

total testing time of 150 minutes, the solution obtained included only items from the quantitative (around 75%) and verbal (around 25%) domains, and none from the English domain. It is worth noting that the English domain currently serves a dual purpose: It is a component of the PET total score, and is used also for placement of students in remedial English classes. Given the above results, which are not altogether surprising in light of predictive validity studies which are conducted routinely at NITE (e.g., Kennet-Cohen, Bronner, & Oren, 1999), the issue of whether to continue to include English as a part of PET or to define it as a separate test for placement purposes only, deserves further consideration.

This last issue, as well as all the other substantive results discussed above, should be treated with reservations. As was mentioned more than once throughout this paper, the application of the proposed method to the context of PET was conducted mainly for illustrative purposes. Thus, for example, for the sake of simplicity, only Hebrew forms of PET were included in the analyses, whereas for predictive validity issues, scores from non-Hebrew forms should also be included. An additional simplification was introduced by averaging and analyzing the results across all university departments, whereas predictive validity considerations should take areas of study into account. Another important limitation of the current application is that PET was treated here as the sole criterion for admission to universities. In practice it is the combination of PET with high school matriculation scores which is used for admission decisions. Thus, the procedure proposed here for validity maximization should actually focus on the incremental validity of PET, beyond the validity of high school achievement, in the prediction of university grades. Another related issue, which limits the practical significance of the results reported here, concerns an assumption made in adjusting the correlations for range restriction. Specifically, it was assumed that PET was the explicit selection variable, whereas, in practice, as was mentioned above, admission decisions are based on a combination of PET and high school matriculation scores. All the above should serve as arguments against deriving any practically significant conclusions from the results reported here. Any further

theoretical or operative conclusions based on empirical results should await additional investigations.

Future applications of the approach proposed here can also provide an opportunity for examining some of the decisions involved in the operationalization of the approach. For example, in the search for the optimal allocation of the testing time, the algorithm for identifying, at each step, the best shift of one unit of testing time from one item type to another proceeded in units of 1 minute. As natural a choice as this may seem, it is clearly an arbitrary one. It might be expected that different results would be obtained when other units of time are adopted (i.e., the best shift of 5 minutes would not necessarily be identical to the end result of 5 shifts of 1 minute). This and several other decisions, which are inevitably involved in the choice of the specific search method to be used, were examined by applying the following alternative methods: the *simplex* method, which is based on a search in the direction of maximal change towards the maximum; *simulated annealing*, which is especially effective in avoiding local maxima; and *differential evolution*, a method which is based on genetic algorithm and is most effective when searching for integer solutions. All these methods were applied using *Mathematica 4* (Wolfram, 1999). All in all, the results obtained from the various methods testified to the acceptability of the method which was ultimately adopted in our study.

The focus of the present approach on a predictive validity perspective should not be interpreted as ignoring or minimizing other considerations pertaining to the test-development procedure. On the contrary, as was demonstrated above, such considerations can be incorporated explicitly by introducing additional constraints into the validity maximization problem. The following are examples of such considerations: a desire to include or exclude items of a certain type on the basis of face validity considerations; a desire to limit the number of items of a certain type because of the expenses involved in constructing them; a need to eliminate a certain type because of adverse effects of coachability; and a wish to maintain a certain number of item types in order to have a more varied and interesting test and to enable a wider

content sampling. All these and other considerations (in the context of PET, for example, the difficulties encountered in translating certain item types is a significant factor; see Allalouf, Hambleton, & Sireci, 1999) can and need to be taken into account in the specification of the constraints. In fact, the proposed approach provides explicit and clear information regarding the price, in terms of predictive validity, to be paid for introducing various constraints on the validity maximization process.

References

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36 (3)*, 185-198.

Attali, Y., & Goldschmidt, C. (1999). *A Rational for the design of the Psychometric Entrance Test*. Research Report No. 268. Jerusalem, Israel: NITE.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer-Verlag.

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.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Gentile, J. R., Kessler, D. K., & Gentile, P. K. (1969). Process of solving analogy items. *Journal of Educational Psychology, 60 (6)*, 494-502.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.

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.}

Kennet-Cohen, T., Bronner, S., & Oren, C. (1999). *The predictive validity of the components of the process of selection of candidates for higher education in Israel*. Research report no. 264. Jerusalem, Israel: NITE.

Moshinsky, A. (2000). *CPET for examinees with disabilities – version 2: results of experiment 2* (Hebrew). Research report no. 276. Jerusalem, Israel: NITE.

Roccas, S., & Moshinsky, A. (in press). Factors affecting the psychometric characteristics of verbal analogies. *Applied Measurement in Education*.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1982). Reasoning, problem solving, and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J. (Ed.)(1984). *Human abilities: An information-processing approach*. San Francisco, CA: Freeman.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

van der Linden, W. J. (Ed.) (1998a). Optimal test assembly. *Applied Psychological Measurement, 22 (3)* [Special issue].

van der Linden, W. J. (1998b). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22,* 195-211.

Whitely, S. E., & Barnes, G. M. (1979). The implications of processing event sequences for theories of analogical reasoning. *Memory and Cognition, 7*, 323-331.

Wolfram, S. (1999). *The Mathematica Book* (4th ed.). Wolfram Media: Champaigne, IL., Cambridge University Press: Cambridge, UK.