Psicothema

Psicothema 2014, Vol. 26, No. 1, 117-126
doi: 10.7334/psicothema2013.257

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2014 Psicothema
www.psicothema.com

Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel

Carmel Oren, Tamar Kennet-Cohen, Elliot Turvall and Avi Allalouf
National Institute for Testing and Evaluation (Israel)

Abstract

Background: The Psychometric Entrance Test (PET), used for admission to higher education in Israel together with the Matriculation (Bagrut), had in the past one general (total) score in which the weights for its domains: Verbal, Quantitative and English, were 2:2:1, respectively. In 2011, two additional total scores were introduced, with different weights for the Verbal and the Quantitative domains. This study compares the predictive validity of the three general scores of PET, and demonstrates validity in terms of utility. **Method:** Sample: 100,863 freshmen students of all Israeli universities over the classes of 2005-2009. Regression weights and correlations of the predictors with FYGPA were computed. Simulations based on these results supplied the utility estimates. **Results:** On average, PET is slightly more predictive than the Bagrut; using them both yields a better tool than either of them alone. Assigning differential weights to the components in the respective schools further improves the validity. **Conclusion:** The introduction of the new general scores of PET is validated by gathering and analyzing evidence based on relations of test scores to other variables. The utility of using the test can be demonstrated in ways different from correlations.

Keywords: validity, standards, predictive validity, higher education, admission.

Resumen

Demostrando la validez de tres puntuaciones generales del PET para predecir el rendimiento en la educación superior en Israel. **Antecedentes:** el *Psychometric Entrance Test* (PET), utilizado para la admisión a la educación superior en Israel junto con la Matriculation (Bagrut), tuvo en el pasado una puntuación general en la que los pesos para sus dominios: Verbal, Cuantitativo e Inglés eran 2:2:1, respectivamente. En 2011 se introdujeron dos puntuaciones totales adicionales, con pesos diferentes para los dominios Verbal y Cuantitativo. Este estudio compara la validez predictiva de las tres puntuaciones generales del PET y demuestra la validez en términos de utilidad. **Método:** muestra: 100.863 estudiantes de primer año de todas las universidades israelíes en los cursos de 2005 a 2009. Se calcularon los coeficientes de regresión y las correlaciones de los predictores con FYGPA. Las simulaciones basadas en estos resultados aportaron la utilidad de las estimaciones. **Resultados:** en promedio, PET es ligeramente más predictivo que el "Bagrut". Asignar pesos diferentes a los componentes en las escuelas respectivas mejora más la validez. **Conclusiones:** la introducción de las nuevas puntuaciones generales del PET es validada mediante la obtención y análisis de evidencia basada en las relaciones de las puntuaciones del test con otras variables. Puede demostrarse la utilidad del uso del test en formas diferentes de las correlaciones.

Palabras clave: validez, standards, validez predictiva, educación superior, admisión.

Test validity is elaborately defined in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999). As the *Standards* describe, validity refers to the degree to which evidence and theory support the interpretations of test scores. The process of validation involves accumulating evidence to provide an empirical basis for the proposed score interpretations.

Although various sources of evidence might be used in evaluating the interpretation of test scores for particular purposes, validity is a unitary concept: the degree to which all the accumulated evidence supports that interpretation. Sources of evidence for validation may include: test content, response processes, test structure and the relationships among its parts, relations to other variables (especially test-criterion relationships) and finally – evidence based on the consequences of testing.

Evidence of the relation of test scores to a relevant criterion may be expressed in various ways, but the fundamental question is always: How accurately do test scores predict criterion performance? The criterion variable is a measure of some attribute or outcome that is of primary interest, as determined by test users. University grades frequently serve as a natural, intuitive external criterion, a special case of "other variables," in relation to which the validity of selection system can be evaluated.

It was testing consequences, namely public criticism of the unitary nature of test scores across different areas of study that, in October 2011, led the Israeli universities to introduce two new general scores, in addition to the existing multi-domain general score of the Psychometric Entrance Test (PET). The new scores

Received: August 27, 2013 • Accepted: October 2, 2013
Corresponding author: Carmel Oren
National Institute for Testing and Evaluation
912600 Jerusalem (Israel)
e-mail: carmel@nite.org.il

are a sciences-oriented general score, and a humanities-oriented general score. The three general scores differ in the weights given to each test domain.

The purpose of this article is to report a current evaluation of the criterion-related evidence in support of the validity of the Israeli system of selection to higher education, including the two new general scores. On this platform, we also present two alternative ways of demonstrating and interpreting the utility of a given test-scores - criterion relationship (predictive validity).

Admission to higher education

Surveys of international practice regarding university admissions (Beller, 1994; Edwards, Coates, & Friedman, 2012; Kellaghan, 1995; McDonald, Newton, Whetton, & Benefield, 2001) reveal a variety of approaches for selecting candidates. One of these approaches, shared by many countries including Israel, is that of using general admissions tests combined with a measure of high school achievement. In the United States, college admissions officers use high school grade point averages (HSGPA) and standardized tests of cognitive ability (SAT or ACT). Selection is conducted either by an explicit formula or, more often, using a holistic approach (Breland, Maxey, Gernand, Cumming, & Trapani, 2002). On the whole, both HSGPA and standardized tests have been shown to have predictive validity in determining a variety of academic performance outcomes. Most of the studies have focused on the prediction of first-year college GPA (Schmitt et al., 2009).

Predictive validity of (cognitive) admission criteria in Israel and in the US

Predictive validity studies investigating the PET in relation to its goals are performed regularly at the Israeli National Institute for Testing & Evaluation (NITE). Most focus on the prediction of first-year university GPA, but from time to time cumulative GPA upon completion of undergraduate studies serves as a predicted criterion as well (c.f., Kennet-Cohen, Bronner, & Oren, 1999b). Beller (1994) described validity coefficients and multiple regression analyses with respect to a previous format of the PET (prior to October 1990). In light of the reported results, it was decided to revise the PET and restructure it. Results with respect to the pre- 2011 format of the PET were analyzed using a meta-analytic approach (Kennet-Cohen, Bronner, & Oren, 1999a) as well as with a more standard approach. The latest reported findings regarding the predictive validity of the components of the process of selection of candidates for higher education in Israel, based on 56,548 observations from 628 departments (Oren, Kennet-Cohen, & Bronner, 2007) show validity coefficients of 0.46, 0.38 and 0.50 for the PET, high school matriculation (Bagrut) mean score and a composite admission score based on equal weighting of the two predictors, respectively (the values are corrected for range restriction). The main conclusion is that the composite score is a better predictor than each of the two components taken separately.

The American SAT is composed, since its revision in 2005, of three sections: Critical Reading (added in 2005), Mathematics and Writing. The SAT score report includes three separate test scores, one for each of the above sections, and there is no official SAT general score. A large-scale national predictive validity study of the SAT was conducted for the 2006 entering freshman cohort in 110 four-year colleges and universities (N = 151,316) (Korbin, Patterson, Shaw, Mattern, & Barbuti, 2008). The correlation of HSGPA and first-year college grade point average (FYGPA) was 0.36, which was higher than the multiple correlation of the SAT (critical reading, math and writing combined) with FYGPA (r = 0.25). The correlations of HSGPA and SAT with FYGPA, corrected for range restriction, were 0.54 and 0.53 respectively. The multiple correlation of HSGPA and all three SAT sections with FYGPA was 0.46 (corrected r = 0.62). Thus, the increment in predictive validity attributable to the SAT is 0.10 and 0.08, in terms of uncorrected and corrected correlations, respectively. The correlations reported above are very similar to those resulting from earlier SAT validity studies (c.f., Bridgeman, McCamley-Jenkins, & Ervin, 2000).

The ACT consists of four tests: English, Mathematics, Reading, and Science. The ACT score report includes a Composite ACT score and four subscores (English, Mathematics, Reading, and Science). The composite score is the average of the four test scores. Correlational evidence regarding the predictive validity of ACT scores concentrates on the ACT composite score; and specifically on the relative predictive validity of ACT composite score and high school average and on the incremental predictive validity of ACT score. Data pertaining to the 2003-2006 entering freshman class years in 192 four-year institutions (N = 120,338) reveal (Sawyer, 2010) that the median correlation of the ACT composite score and high school average with FYGPA were 0.48 and 0.41 respectively; and that the median multiple correlation of both predictors with FYGPA was 0.54 (these values are not corrected for range restriction). This finding - that high school average is a better predictor of first-year college GPA than the ACT composite score, but the ACT composite score has incremental predictive validity - was reported in previous studies as well (ACT, 1999, 2008).

It should be noted that much of the validity evidence reported for ACT (c.f., ACT, 2007; Sawyer, 2010) concentrated on decision-based statistics (such as accuracy rate and success rate) instead of on the predictive strength of the selection variables (such as measured by correlations). The results from this perspective suggest that high school GPA is more useful than the ACT composite score in situations involving low selectivity in admissions and minimal-to-average academic performance in college as the criterion of success. In contrast, the ACT composite score is more useful than high school GPA in situations involving high selectivity and high academic performance as the criterion of success. In nearly all contexts, test scores have incremental utility beyond high school GPA (Sawyer, 2010).

To sum up, ACT validity research focuses on the ACT composite score. With respect to the SAT, no official composite score is reported, and validity studies focus on the multiple correlations of the scores in the three SAT sections with the FYGPA criterion. Thus, in both contexts there is no consideration of alternative definitions of the composite score, as proposed with respect to PET. The results of the vast majority of predictive validity studies in the US indicate that HSGPA is usually better than admission test scores in predicting FYGPA, although test scores have incremental predictive validity.

Computing general scores based on subscores

Computing general scores from subscores is common for selection purposes. In educational or other contexts of admissions, the selection is usually performed on the basis of one general, final score that serves to put all the candidates on the same scale. Calculating a general score from subscores can be done by several methods. If there is no theory regarding the importance or relevant psychometric characteristic of each subtest, the subscores should be weighted equally (Raju et al., 1997). If this is not the case, and the subtests differ by importance, reliability, time allocation or face validity, different weights can be used. An empirical approach, if there is a criterion for success, leads to the use of a linear (or other) regression to find the optimal weights for prediction. Different weights for subscores may also be used when the prediction is used for different purposes, even in the same institution, for example, different schools in a university (we use the term "school" to denote a group of departments dealing with closely related areas of study in a university or college). It is widely agreed nowadays that for the sake of transparency, the weights should be made public (AERA et al., 1999).

The background of creating the two additional general scores for the PET

PET is a high-stakes multiple-choice test developed and administered by NITE. It is used for admission to universities and other institutions of higher education in Israel (Beller, 1994). The test consists of three subtests/domains: Verbal Reasoning (V), Quantitative Reasoning (Q); and English as a Foreign Language (E). Each domain comprises two test sections.

To give examinees the best chance of demonstrating their ability in each of the three domains being assessed, the PET is translated and adapted into five languages. The test in all its variations is administered to about 75,000 examinees annually, in five different exam dates. Each year, about 50,000 examinees are tested in Hebrew, 20,000 in Arabic; and a total of 5,000 in Russian, English, French and Spanish.

The PET general scores

Between October 1990 and July 2011 one General Score was calculated: A Multi-Domain General Score (TGE) in which the weight of the scores in the Verbal Reasoning and Quantitative Reasoning domains is double the weight of the score in the English domain (Allalouf, 1999). As of October 2011, NITE added two General Scores: (1) A Humanities-oriented score (THU), in which the score in the Verbal Reasoning domain is three times the weight of each of the other two scores; and (2) A Sciences-oriented score (TSC), in which the score in the Quantitative Reasoning domain is three times the weight of each of the other two scores. The additional General Scores were introduced for both potential predictive validity and face validity improvement: the extent to which each of the PET domains (Verbal Reasoning, Quantitative Reasoning; and English) is relevant to academic studies depends on the course of study being pursued. In sciences-oriented departments (mathematics and physics, for example) the Quantitative Reasoning domain is much more relevant than it is in humanities-oriented departments such as literature or history. For other departments or schools, such as law, the Verbal Reasoning domain seems much more relevant. All the institutions that receive PET scores from NITE can be provided, upon request, with the information needed to use the new scores.

Method

Population

The data for this study were supplied by all first-year students of all (six) Israeli universities over the five academic years 2005/06-2009/10. The universities are: Ben-Gurion University of the Negev, Bar-Ilan University, the University of Haifa, the Hebrew University of Jerusalem, Technion - Israel Institute of Technology and Tel Aviv University.

Variables

Predictors. We investigated several predictors in this study and also studied different weighting strategies for these predictors. The list of predictors and weighting strategies follows.

High School Matriculation tests:

1. B - Bagrut - Matriculation test-battery mean score. The Matriculation Certificate is based on a series of national, supposedly objective, subject-matter test scores, backed by school achievement tests. In reality, the abundance in test forms and school levels jeopardizes the comparability, the objectivity, and thus the reliability and the predictive validity of this score as a fair high-stakes means of selection to higher education.

Three Psychometric Entrance Test component subscores:

2. V - Verbal reasoning.
3. Q - Quantitative reasoning.
4. E - English as a foreign language.

Three PET Total Scores, differing in the relative weights of the three components [presented in square parentheses]:

5. TGE - General (multi-domain) psychometric score of PET [2V, 2Q, E].
6. THU - Humanities-oriented score [3V, Q, E].
7. TSC - Sciences-oriented score [V, 3Q, E].

Three Composite Admission Scores comprised of PET Total Scores and Bagrut (B), with equal weights:

8. CGE [TGE, B].
9. CHU [THU, B].
10. CSC [TSC, B].

Criterion. The criterion for all analyses was First Year Grade Point Average (FY) in one of the aforementioned Israeli universities. Since the grading standards and policies may vary considerably across universities and among departments within universities, we refer to the relative value of the criterion scale within departments, as described in the next paragraph "Units of analysis."

The three Composite Admission Scores were defined as equally-weighted PET Total Scores and B score at candidate level at each university. The composite admission score was constructed in the following way: Means and standard deviations of B and the three PET general scores were computed at the level of the candidates, for the academic years 1992/3 and 1993/4, within each university.

For each university, these statistics were averaged across the two academic years, weighted by the number of candidates. The composite admission score was computed as the sum of each PET general score and B score, standardized according to the statistics above.

Units of analysis

The statistics were computed within institutions, departments and year of study, wherever at least 20 students had data on all predictors and the criterion, and are presented as weighted averages by the department's size, within School: **Hum**anities, **Soc**ial-**V**erbal (education, psychology and political science), **Law**, **Soc**ial-**Q**uantitative (statistics and economy), **Nat**ural sciences (biology, mathematics and physics), **Eng**ineering (engineering and architecture), **Med**icine (medicine, dentistry and pharmacy); and **Para**-medical professions (nursing, occupational therapy, physiotherapy; and speech therapy).

Analyses

The following statistics are shown as measures of predictive validity: Multiple linear regression weights of the predictors; simple and multiple correlations (corrected for restriction of range) of the predictors with the criterion, and the alternative terms of the results. (Means, s.d.'s and frequencies of the research variables are presented in tables 1-3.)

The correction for restriction of range was based on the assumption that the selection was done by CGE (TGE+B) score. The other predictors were only exposed indirectly to the selection process. The model for correction in this situation was described by Gulliksen (1950) for the bivariate case and for the three-variable case (chapters 11 and 12). As an estimate of the population's standard deviation of CGE score, we used the mean of the standard deviations of CGE score among candidates to each department over two academic years (1992/3 and 1993/4), weighted by the number of candidates in each department.

Results

Descriptive statistics

The frequencies, the means and the standard deviations of all the variables in the study are presented in tables 1, 2 and 3.

Table 1
Number of students and departments (in parentheses) by school and study year

All	Para	Med	Eng	Nat	Soc-Q	Law	Soc-V	Year	Hum
19687	1140	672	2770	3387	2109	857	5979	2005	2773
(227)	(15)	(9)	(34)	(48)	(16)	(4)	(43)		(58)
20719	1307	752	3030	3462	2237	867	6404	2006	2660
(216)	(15)	(11)	(34)	(45)	(14)	(4)	(43)		(50)
19710	1194	753	3053	3264	2117	782	6084	2007	2463
(217)	(15)	(11)	(34)	(45)	(14)	(4)	(42)		(52)
20579	1261	782	3392	3422	2226	854	6143	2008	2499
(215)	(15)	(12)	(33)	(44)	(14)	(4)	(43)		(50)
20168	1279	924	3282	3314	2147	778	5906	2009	2538
(220)	(15)	(14)	(33)	(46)	(14)	(4)	(43)		(51)
100863	6181	3883	15527	16849	10836	4138	30516	All	12933
(1095)	(75)	(57)	(168)	(228)	(72)	(20)	(214)		(261)

Table 2
Means of the research variables by school

School	FY	CGE	CHU	CSC	B	TGE	THU	TSC	V	Q	E
Hum	84	55	56	54	97	602	608	597	119	114	124
Soc-V	83	56	56	55	97	583	585	581	115	114	115
Law	82	67	67	66	105	689	692	687	134	132	134
Soc-Q	80	64	64	64	103	676	671	680	129	133	130
Nat	76	63	63	63	104	659	654	665	126	130	128
Eng	78	64	63	64	105	671	661	681	126	135	129
Med	84	69	69	69	109	707	706	708	136	137	136
Para	82	58	58	58	100	600	600	600	118	118	114
All	81	60	60	60	101	632	631	633	122	124	124

Table 3
Standard deviations of the research variables by school

School	FY	CGE	CHU	CSC	B	TGE	THU	TSC	V	Q	E
Hum	9.1	8.8	8.9	8.7	8.1	77	79	78	16	16	17
Soc-V	8.3	7.1	7.3	7.1	7.5	65	68	66	14	14	17
Law	8.0	4.1	4.1	4.2	5.5	38	40	41	9	9	12
Soc-Q	9.8	4.5	4.7	4.5	6.1	39	43	39	10	9	13
Nat	13.3	5.5	5.8	5.4	6.1	53	58	52	13	10	15
Eng	9.4	4.8	5.1	4.6	5.5	44	51	41	12	8	14
Med	7.2	4.1	4.2	4.2	5.0	34	37	35	8	8	11
Para	6.9	6.1	6.2	6.1	7.5	48	52	50	12	11	17
All	9.6	6.3	6.5	6.2	6.8	57	61	57	13	12	15

Regression weights

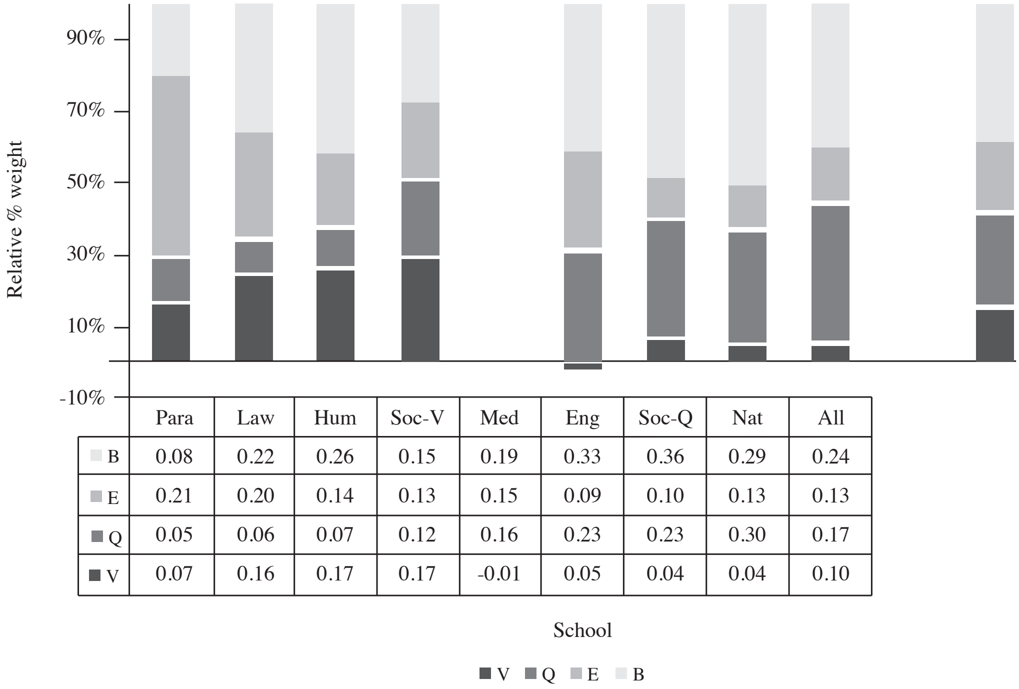
A four-variable (V, Q, E, B) model of multiple linear regression, reconstructed from all the intercorrelations (corrected for restriction of range), yielded for each school the mean standardized (β) coefficients of the admission components in predicting FY. These weights are presented in the table at the bottom of Figure 1. The proportions within the columns of the graph represent the relative weight of each of the four predictors, summing up to 100% within each school.

The results indicate that the schools can be grouped into two bulks: the exact/sciences-oriented schools, in which Q weights by far (at least four-times) more than V: Nat, Soc-Q, Eng, Med; and the humanities-oriented schools, in which V (and E) weight considerably-to-moderately more than Q: Soc-V, Hum, Law, Para. The size of the improved predictive validity will be shown in the next section.

Correlations

The validity coefficients, corrected for the restriction of range of the predictors, are presented in Table 4 (The raw (observed) correlations are presented in Table 5). The presented multiple R, based on a four-variable model regression, is reconstructed from the matrix of corrected validity coefficients.

The highest correlation within each triad of predictors is highlighted in the table. In accordance with the regression weight results, the following pattern of results is seen among the correlations:



	Para	Law	Hum	Soc-V	Med	Eng	Soc-Q	Nat	All
B	0.08	0.22	0.26	0.15	0.19	0.33	0.36	0.29	0.24
E	0.21	0.20	0.14	0.13	0.15	0.09	0.10	0.13	0.13
Q	0.05	0.06	0.07	0.12	0.16	0.23	0.23	0.30	0.17
V	0.07	0.16	0.17	0.17	-0.01	0.05	0.04	0.04	0.10

Figure 1. *Relative weights (standardized linear regression β coefficients) of three PET's components: V, Q, E; and Bagrut (B), in predicting FY, by school*

Table 4
Corrected correlations* of predictors with FY, by school

	Mult. R	**Composite score**				**PET total score**					
School	**(V,Q,E,B)**	**CGE**	**CHU**	**CSC**	**B**	**TGE**	**THU**	**TSC**	**V**	**Q**	**E**
Law	54	44	45	43	37	45	46	43	39	34	38
Hum	56	48	49	47	38	43	43	41	39	34	34
Soc-V	49	41	41	40	27	42	42	40	37	33	33
Para	42	19	20	18	12	24	24	22	18	13	24
Med	51	35	34	36	30	33	30	35	24	31	24
Soc-Q	59	53	52	54	47	43	39	44	31	39	29
Nat	63	57	55	58	45	52	48	54	42	50	38
Eng	58	53	52	54	46	45	43	47	37	42	31
All	54	46	46	46	36	43	41	43	36	37	32

** Decimal point omitted*

Table 5
Raw (observed) correlations* of the predictors with FY, by school

School	**CGE**	**CHU**	**CSC**	**B**	**TGE**	**THU**	**TSC**	**V**	**Q**	**E**
Hum	48	49	47	38	43	43	41	39	33	34
Soc-V	33	33	32	16	34	34	33	30	25	26
Law	23	25	21	10	24	25	21	20	11	23
Soc-Q	32	29	34	22	22	17	25	10	23	13
Nat	41	38	43	24	34	29	37	22	35	22
Eng	33	29	35	23	22	18	25	12	24	12
Med	20	18	22	11	18	14	20	06	17	15
Para	11	12	10	00	17	18	15	11	06	21
All	34	33	34	21	30	28	30	22	25	22

* Decimal point omitted

In the humanities-oriented schools, the highest validities are those where the verbal component received augmented weight in the total PET score, thus CHU >= CGE. In the exact / sciences-oriented schools, the highest validities were reached by augmenting the weight of the quantitative component Q in PET Total Score, resulting in CSC > CGE.

Overall, the mean validity of CGE is 0.46 - higher than the validity of any PET general score (0.43), which in turn is higher than the validity of B (0.36). Using optimal weights within units of analysis would result in an average multiple correlation of 0.54. Within schools, the combination of B and a PET general score yields most of the time higher validities than each of them used alone. The exceptions are the schools of Law, Soc-V and Para, in which the use of B and PET with equal weights does not seem to contribute incremental validity to PET taken alone.

Validity interpretation

In this section, we move beyond the issue of comparing the validity of the three general scores, and discuss predictive validity in general. To interpret predictive validity results in a more comprehensible way, we created a demonstration that shows the benefit of using a selection system with characteristics identical to ours. We show the benefit in two sets of terms: The gain in achievement on the criterion scale (in standardized scores); and the gain in proportion of correct placements of candidates into their actual ability groups.

Simulation

Our example is based on the results of a simulation. This analysis was run with identical validity coefficients to the ones presented in Table 4, for the school Nat (Natural Sciences).

The need for simulation stems from the fact that we want to show the results with “true” validities, while the empirical study group is only a truncated sample - of accepted students. For that matter we simulated the whole population of candidates. The truncation, often named restriction of range, was corrected statistically for the predictors prior to the simulation. The criterion was not directly corrected for its presumed restriction of range, but within the simulation, some correction was made by generating all three variables involved with full normal distributions.

The simulation was run on 10,000 virtual observations. (Simulations on 10,000 give very close results to the true situation, but still produce some random error. This can be seen when comparing the opposite cells on the joint distribution - the trinomial distribution should produce symmetric results). The simulated selection system mimics reality: it encompasses two predictors, say B and TGE, with validities of 0.45 and 0.52, respectively, in predicting a third variable, say FY; and a correlation of 0.44 between the two predictors. We used standardized (0, 1) scores for all three variables, in order to free ourselves from specific score scales, and also to obtain easily interpreted results in standardized difference terms.

All three variables - the two predictors and the criterion - were grouped into quartiles. The 25 percent of the observations of any given variable quartile do not consist, of course, of the same observations as the respective quartile of the other two variables. (The higher they correlate with one another – the more they would overlap.) The results are shown in 4 by 4 cross-tabulations in Table 6.

Validity and score gain

The mean standard score of FY within each combination of the two predictor quartiles, and the marginal means (within each predictor quartile, independently of the other) are shown in table 6.

From this table, many indices of the benefit of predictive validity can be derived and interpreted in terms of standardized difference. For instance, the d (standardized difference) values of mean FY between quartile combinations of the predictors are presented in Table 7. The compared cells are highlighted, and explained in Table 7 footnotes.

The standardized differences show for each validity value the gain in the criterion score between the groups of ability measured by the predictors. They are larger as the validity grows and they still grow considerably when using both predictors, as we compare combinations of quartiles. The standardized difference is considerable even between adjacent cells.

As to the visualization of all three variables from Table 7 simultaneously: Figure 2 presents the mean FY score within each quartile of B, for each level of TGE. Figure 2 demonstrates the significant incremental contribution of TGE to the selection process, beyond the effect of B. For instance, within the upper (Q4) B quartile, FY mean score grows half a s.d. higher when moving from Q3 to Q4 on TGE. The opposite, namely the incremental validity reached by adding B to TGE, is also evident but to a lesser extent, since B is slightly less valid than TGE.

Table 6
Mean standard score (0, 1) of FY in quartiles of TGE and B

	TGE Quartile				
B Quartile	Q1	Q2	Q3	Q4	ALL
Q1	-0.90	-0.47	-0.17	0.19	-0.55
Q2	-0.62	-0.24	0.06	0.38	-0.16
Q3	-0.37	-0.04	0.19	0.58	0.13
Q4	-0.14	0.19	0.48	0.92	0.58
ALL	-0.66	-0.17	0.17	0.67	0.00

Table 7
Validity and interquartile differences in criterion scores

Predictor	Validity	D (Q4-Q1)	D (Q4-Q3)
B*	0.45	1.13	0.45
TGE*	0.52	1.33	0.50
B+TGE**	0.57	1.82	0.73

** The compared cells are the extreme ones in the margins.*
*** The compared cells here lie on the diagonal of the table: The two extreme ones and the two adjacent; and they are not between quartiles, but rather between same-ability-on-predictors groups*

Validity and the proportion of successful placements

Predictive validity can also be translated into percent of students successfully placed into their expected group of criterion ability. The more valid the predictor, the more high-FY students will be found among the high-ability students as measured by the predictor. Low validity will be manifested in random dispersion. The same aforementioned simulation characteristics were used for the following data. Table 8 shows the proportion of excelling students, which belong to the upper quartile on FY, within each combination of quartiles of the two predictors, B and TGE. The base-rate for comparison, in a random (zero-validity) selection situation, is 25%.

The margins of Table 8 show that the higher the level of each of the two predictors, the larger the proportion of excelling (Q4) FY students is found: for instance, moving from Q3 to Q4 on B increases the FY Q4 percent from 28 to 46; and from 28 to 50 percent when moving from Q3 to Q4 on TGE. The simultaneous three-variable picture is displayed in figure 3.

Figure 3 shows that within each level of B, the percent of upper quartile FY students grows dramatically as their TGE level grows and vice-versa, but to a lesser extent, within each quartile of TGE, as their level of B grows. The translation of validity into distributional terms can serve to derive measures of correct-placements, as a function of use of a combination of predictors with given predictive validity.

Conclusion and discussion

The selection system used by the Israeli universities and the scores reported to clients (the admissions offices), have evolved both on academic-criterion-related, and on consequence-related grounds.

Table 8
Percentage of Highest Quartile (Q4) FY students in Quartiles of TGE and B

	TGE quartile				
B Quartile	Q1	Q2	Q3	Q4	ALL
Q1	4	7	14	30	9
Q2	6	13	23	36	18
Q3	10	21	27	47	28
Q4	14	27	41	61	46
ALL	6	16	28	50	25

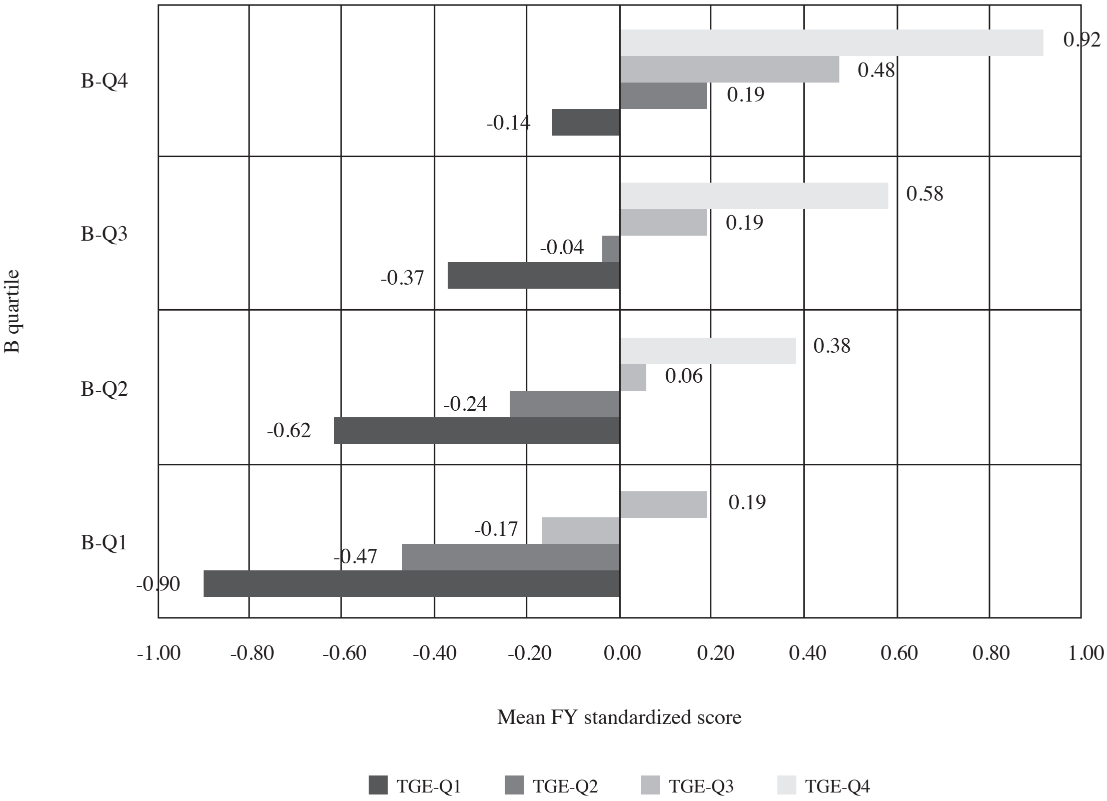


***Figure 2.** FY standardized scores in quartiles of B and TGE*

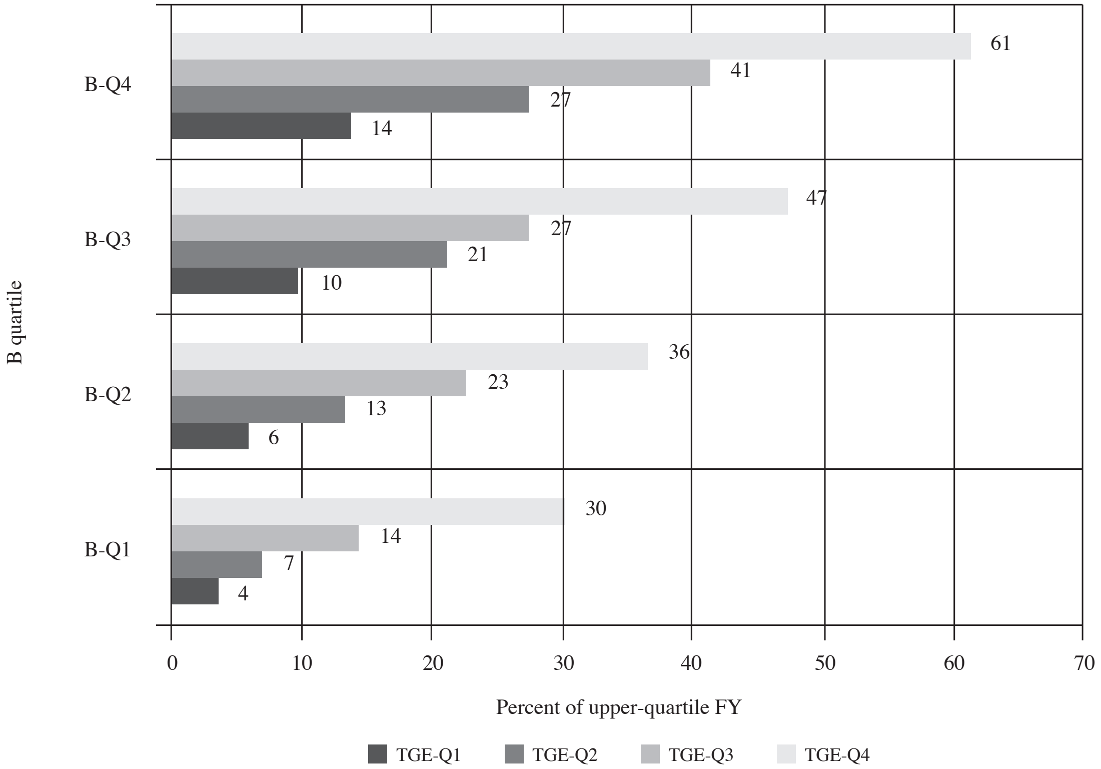


***Figure 3.** Percent of Upper Quartile (Q4) FY students in Quartiles of B and TGE*

On the basis of empirical evidence, which is in accordance with intuitive and face-value considerations, NITE introduced in 2011 the two new general (Total) scores for PET: THU – Humanities-oriented score [3V, Q, E]; and TSC – Sciences-oriented score [V, 3Q, E]. These two new general scores are offered as "ready-made" proxies, to help admissions officers gain, along with simplicity of use, both face validity and improved predictive validity.

The current big validity-picture is the following: On average, PET is slightly more predictive than the Bagrut average score: 0.43 and 0.36, respectively; using them both weighted equally, yields in most cases a composite measure which is more predictive than either of its components alone: 0.46. Attention should be paid to departments in the schools of Para-Medical studies, Law and Social Sciences (verbal), where adding Bagrut to PET with equal weights is lowering the validity of the composite score compared to PET alone. Using optimal regression weights of all four components the three PET subscores and the Bagrut, calculated within departments, would produce an average validity of 0.54.

These results are roughly similar to those reported for the SAT, the reported multiple R of which is 0.46 (corrected r = 0.62). (It should be noted though, that our correction for range restriction of the PET is more conservative than the SAT's: It was done on the basis of the actual candidates' parameters, while the correction of SAT – on the basis of all College-Bound-Seniors cohort parameters, who are theoretically, in a world without selection, the potential candidates.)

The introduction of the two new General Scores of PET, the Sciences-oriented score and the Humanities-oriented score, seem to have delivered the expected results. Both the predictive and the face-validity (Karelitz, 2013) benefit from the change, although the increment in predictive validity is small: about one point (1%) of Pearson correlation coefficient. The effect is further diluted when adding the Bagrut to PET with equal weights. But as long as the prediction is improved and the rationale of a more efficient and relevant-to-context use of the existing selection components is satisfied – the reform appears to be justified.

Despite all this, it should be emphasized that, as in the past, it is still the case today that the institutions of higher education in Israel can assign each of the PET domains a weight different than that assigned by NITE in the General Scores it reports. All decisions regarding the admission of candidates are the sole responsibility of the institutions of higher education.

In addition to the predictive validity picture reported, we present a demonstration of the meaning of a given predictive validity result, in more concrete, day-to-day terms of utility. We offer two conceptual ways to interpret predictive validity: In terms of achievement gains and in terms of gain in proportion of correct placements. The idea is to divide the whole range of each variable into quartiles and to cross-tabulate them. The inter-quartile differences on scores or on proportion of a given group of interest can help interpret the correlation of the validity into a more comprehensible language. Tables of these equivalents can be calculated for the whole range of validities; and for different variations of combining predictors in a selection system. The importance of supplying validity data to the public in a communicative manner cannot be overrated.

High-stakes tests like PET, used in an economically and culturally diverse society, need a strong body of empirical evidence to justify their use. The argumentation of content relevance is not enough, as is the claim of generalizability of the test's general-score predictive power across domains of study. This study is an example of using the relation of test scores with an external variable – first-year university grades – to interpret the correlations in terms of gains in academic achievements, and in better placement of candidates. Furthermore, we show that by using differential weighting of test components in different areas of study, we achieve a slight improvement in the prediction precision and also advance public acceptance of the test's proposed interpretative claim.

References

ACT (2007). *The ACT technical manual*. Iowa City, IA: ACT, Inc.

ACT (1999). *Prediction research summary tables* (Available from ACT, Inc., P. O. Box 168, Iowa City, Iowa 52243).

ACT (2008). *Updated validity statistics for ACT Prediction Service* (Available from ACT, Inc., P. O. Box 168, Iowa City, Iowa 52243).

Allalouf, A. (1999) Scoring and equating at the National Institute for Testing & Evaluation. *Research Report No. 269*. Jerusalem: NITE.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

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

Breland, H., Maxey, J., Gernand, R., Cumming, T., & Trapani, C. (2002). *Trends in college admission 2000. A report of a survey of undergraduate admissions policies, practices, and procedures*. ACT., Inc., Association for Institutional Research, The College Board, Educational Testing Service, The National Association for College Admission Counseling.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). Predictions of freshman grade-point average from the revised and re-centered SAT I: Reasoning Test *(College Board Report No. 2000-1)*. New York: College Entrance Examination Board.

Edwards, D., Coates, H.B., & Friedman, T. (2012). A survey of international practice in university admissions testing. *Journal of Higher Education Management and Policy, 24*(1), 87-104.

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

Karelitz, T.M. (2013). Using public opinion to inform the validation of Test Scores. *Research Report no. 387. Jerusalem:* NITE.

Kellaghan, T. (Ed.) (1995). *Admission to Higher Education: Issues and Practice*. Princeton, NJ: International Association for Educational Assessment.

Kennet-Cohen, T., Bronner, S., & Oren, C. (1999a). A meta-analysis of the predictive validity of the selection process used by universities in Israel. *Megamot, 40*(1), 54-71, in Hebrew.

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

Korbin, J.L., Patterson, B.F., Shaw, E.J., Mattern, K.D., & Barbuti, S.M. (2008). *Validity of the SAT for Predicting First-Year College Grade Point Average (Research Report No. 2008-5)*. New York: College Board.

McDonald, A., Newton, P., Whetton, C., & Benefield, P. (2001). *Aptitude testing for university entrance: A literature review*. Slough: NFER.

Oren, C., Kennet-Cohen, T., & Bronner, S. (2007). Grouped data regarding the validity of the university selection tools in predicting freshman

GPA (Cohorts 2002-2004)]. *Research Report no. 342. Jerusalem:* NITE, in Hebrew

Raju, N.S., Bilgic, R., Edwards, J.E., & Fleer, P.F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement, 21*(4), 291-305.

Sawyer, R. (2010). Usefulness of high school average and ACT scores in making college admission decisions (*ACT Research Report No. 2010-2)*. Iowa City, IA: ACT, Inc.

Schmitt, N., Keeney, J., Oswald, F.L., Pleskac, T., Quinn, A., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and non-cognitive predictors and the impact of demographic status on admitted students. *Journal of Applied Psychology, 94*, 1479-1497.