מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
المركز القطري للامتحانات والتّقييم
מיסודן של האוניברסיטאות בישראל

General Ability or Distinct Scholastic Aptitudes? A Multidimensional Validity Analysis of the Psychometric Higher-Education Entrance Test

Report
RR
16-08

September 2016

Dvir Kleper ❙ Noa Saka

NITE REPORT RR-16-08

ISBN:978-965-502-200-1

General Ability or Distinct Scholastic Aptitudes? A Multidimensional Validity Analysis of the Psychometric Higher-Education Entrance Test

Dvir Kleper ❙ Noa Saka

September 2016

Table of Contents

List of Tables

List of Figures

Abstract

The present study explored the construct validity of the Psychometric Entrance Test (PET) for higher education in Israel, as represented by the factorial structure of the scholastic aptitudes it measures, and focused on whether the test presents a single measure of overall ability or a measure of the fields of knowledge that are being tested. In Study 1, we used Exploratory Factor Analysis to generate hypotheses regarding the factorial structure of the test. In Study 2, Confirmatory Factor Analysis was carried out to compare competing models that were constructed based on theoretical considerations and the results of Study 1. The findings indicated that a two-layered hierarchical model, encompassing both a *General Ability* factor and three scholastic domain-specific factors (*Verbal Reasoning*, *Quantitative Reasoning*, and *English*), showed the best fit. Within the framework of the CFA, several statistical procedures were applied to assess reliability (indicator and complexity) and validity (convergent and divergent).

Introduction

High-stakes tests have long been used by education systems to make decisions regarding applicants, including acceptance to or rejection by institutions of higher education (Amrein & Berliner, 2002). A high-stakes test is generally defined as one that carries a great deal of weight, and the results of which are likely to have a significant impact – for better or for worse – on a specific individual or organization (Zwick 2007). Selection tests for higher education are generally used to predict a candidate's academic success and to rank applicants, however each institution is free to decide how much weight these tests will have in its admissions process, something which places a great deal of pressure on candidates to do well. There is tremendous controversy regarding the need for these tests, their validity and reliability, and the weight they should be given.

The stakes rise as the potential results place the examinee in a position of greater risk or uncertainty. Inasmuch as these tests play an important role in university admissions, professional training (law boards), and job placement, they must have the psychometric properties that will ensure their ability to effectively predict those criteria that they are designed to predict. For this reason, the psychometric quality of the tests, and a description of their psychometric properties, is crucial.

The Internal Structure of Scholastic Aptitude Tests

Numerous studies have focused on issues pertaining to the validity and reliability of such tests (for a review, see Zwick 2007). At the same time, a sizeable minority of studies have examined the internal structure of the constructs measured by aptitude tests and the individual components that comprise them. Some studies looked at the internal structure of test items or scoring scales as part of a broader validity assessment or to compare functionality between groups (for example, Cahalan-Laitusis, Cook & Aicher, 2004; Cook, Eignor, Steinberg, Sawaki & Cline, 2014; Huynh & Barton, 2006; Steinberg, Cline & Sawaki, 2011).

Only a few studies have investigated the internal structure of high-stakes tests used for university admissions. Some of them did so in an attempt to examine whether these tests present a single measure of overall ability or a measure of the fields of knowledge that are being tested (Koening, Frey & Detterman, 2008). In general, these studies yielded conflicting results regarding the question of whether the tests in fact

assess a single factor (presumably, general ability) or whether it is possible to identify sub-domains of scholastic aptitudes that do not measure a single general ability but are also influenced by domain-specific knowledge. For example, Rock, Bennett, Kaplan & Jirele (1988) looked at the internal structure of the US-based SAT and GRE. They found that the SAT consisted of two factors representing the content domains (verbal and quantitative) that made up the test at the time. The GRE comprised three factors (verbal reasoning, quantitative reasoning, and analytical reasoning).

In contrast, other studies found very high correlations between *g*, for general ability, and both the SAT and the ACT (Frey & Detterman, 2004; Koening, et al., 2008). These researchers concluded that, as the SAT and ACT were strongly *g* loaded, they actually measured a single construct, namely general intelligence.

To further investigate these contradictions, Coyle and Pillow (2008) conducted a path analysis and found that, while the SAT and ACT were highly loaded on the *g* factor, both tests predicted college GPA after controlling for *g*. These findings suggest that while both tests are highly related to one general intelligence factor, they also consist of other non-*g* factors that contribute to their predictive ability.

Thus, one of the objectives of the present study was to further examine this question by assessing the internal structure of a test used for admissions to higher education in Israel – the Psychometric Entrance Test (PET), which has a format and goals similar to those of tests used in the US (e.g., SAT, ACT, GRE) – and by determining whether the factors being tested reflect a single general ability factor or ability in the various academic domains (Verbal Reasoning, Quantitative Reasoning, and English) measured by the test.

The Psychometric Entrance Test

The Psychometric Entrance Test (PET) is a tool used to select students for undergraduate programs in Israel and to predict their chances of academic success. The PET, considered an objective and standardized test that has good predictive validity, is used by numerous institutions of higher education to rank applicants to various departments. The PET can be used to rank candidates on a single, uniform scale and, compared to other selection tools, is influenced less by the student's background or other subjective variables (Kleper, Allaluf, Turvall, Oren, & Pronton, 2015).

In most cases, the institutions of higher education calculate a *Composite Score* that is made up of an overall high school matriculation score as well as the PET score (usually equally weighted). Applicants are accepted to institutions of higher education based on this composite score, in addition to other criteria that may be set by certain departments. In some cases, however, the PET provides a second chance to examinees with good potential who, for various reasons, did not fully demonstrate their capabilities during high school. Moreover, because it is translated into several languages, the PET offers a uniform scale for examinees whose mother tongue is not Hebrew or who do not have an Israeli matriculation certificate. In light of these characteristics, it is critical that ongoing research on the test's validity and psychometric properties be conducted so that it can be used by institutions of higher education.

The PET is a moderately speeded test, consisting of three sections, each one corresponding to a scholastic domain – Verbal Reasoning, Quantitative Reasoning, and English. Over the years, PET's format and structure have undergone significant changes in order to bolster face validity and remove items that make a relatively small contribution to its predictive capability, or that fall outside of test parameters (i.e., items that require memorization of vocabulary words). Recently, a writing task component, accounting for 25% of the score in the Verbal Reasoning domain, was added to the PET.

Many studies have shown that the PET has good predictive validity, that is, most examinees who scored higher on the PET demonstrate greater success in their undergraduate studies than examinees who scored lower (Beller, 2001; Kennet-Cohen, Bronner, & Oren, 1999; Kleper & Turvall, 2016; Kleper, Turvall, & Oren, 2014; Oren, Kennet-Cohen, Turvall & Allalouf, 2014). Moreover, it has been found that of all the possible combinations of existing selection tools, the combination of high school matriculation score and the PET score has the best predictive capability (Beyth-Marom et al., 1998; Kennet-Cohen, Oren, Turvall, & Cohen, 2013).

Only two studies had previously been conducted into the internal structure of the PET, and both of those were based on very old test formats. Budesco (1985) investigated the structure of the PET in its original format (when it consisted of five domains: shapes, mathematics, verbal understanding, English, and general knowledge). His study found two factors – knowledge (encompassing general knowledge and

English) and quantitative reasoning (encompassing mathematics and shapes); the verbal reasoning domain was linked to both factors. Beller (1990) conducted another study that used three methods of structural analysis: Exploratory Factory Analysis, ADDTREE (Sattath & Tversky, 1977), and Smallest Space Analysis (Guttman, 1968). This study also identified two factors – knowledge and problem solving. Since then, no studies have been carried out on the test's internal structure despite the fact that it has undergone several alterations. Therefore, an evaluation of the internal structure of the test is a necessary step towards ensuring its validity.

Goals of the Present Study

The main goal of the present study was to characterize the internal structure of the scholastic aptitude assessed by higher education selection tests, in this case the PET. Specifically, we examined whether this structure is better described as a single-ability, one-factor construct (therefore justifying the use of only one score) or whether it is better conceptualized as a thee-factor construct, corresponding to the three academic test domains (therefore justifying the use of three domain-specific scores and various combinations thereof).

Several methods may be used to analyze the internal structure of test items and scales. Common ones are using cluster analysis or exploratory factor analysis to analyze matrices of the intercorrelations between variables and the production of general score indexes (Beller, 1990). Over the past few years, new models have been developed that allow us to verify, using Confirmatory Factor Analysis (Bentler, 1990, 1992), *a priori* assumptions vis-à-vis the internal structure of scales and factors.

Thus, in the present study we conducted two sets of analyses. In Study 1, we employed Exploratory Factor Analysis (EFA) on the score indexes yielded by the test to uncover possible underlying structures. In Study 2, we used Confirmatory Factor Analysis to test several competing models for the underlying structure, based both on the results of the EFA and on other theoretical considerations, including the assumption that the empirical structure of the intercorrelation matrices among the test's various scales indeed give rise to a structure comprised of three factors that fit well with the data.

Study 1

Overview

Construct validity is a central issue in assessing the psychometric properties of tests, when inferences must be made concerning unobservable or latent variables. Factor analysis is an important tool for questions of validity and the measurement of psychological constructs (Nunnally, 1978) and is one of the most commonly used procedures in the development and evaluation of psychological measures. Factor analysis is particularly useful with multi-item inventories designed to measure multifaceted constructs. Therefore, in Study 1 we used scale-level Exploratory Factor Analysis (EFA) to generate hypotheses regarding the factor structure of the test, hypotheses that were subsequently subjected to further investigation by Confirmatory Factor Analysis (CFA) in Study 2.

Method

Sample: To conduct the EFA, we selected one PET administration. The test version that was chosen was administered in December 2012 (in Hebrew) to 4,799 examinees. Males comprised 38% of the sample and females 62%; average age was 21.7 years (standard deviation of 2.7, median 21.6). Native Israelis made up 82% of the sample. About 2.5% of examinees reported that their socioeconomic status was very much higher than average, 22.6% much higher than average, 41% slightly above average, 20.3% slightly below average, 10.1% much below average, and 3.6% very much below average. Statistical comparisons indicated that this sample was representative of the population of examinees during the December test administration over the previous 10 years.

Materials

Test Structure and Components: The PET comprises nine sections, each of which represents one of three domains: Verbal Reasoning, Quantitative Reasoning, or English. The first section of the test is the writing task, which is part of the Verbal Reasoning domain. The remaining eight sections consist of multiple-choice questions, for each of which the examinee is required to select the correct answer from among four possibilities. These sections, which do not appear in any set order, will be referred

to hereafter as multiple-choice sections. The number of questions and the time allotted for solving them are noted at the beginning of each section.

The multiple-choice sections in each domain contain several question types. Questions of each type are grouped together and arranged in order of increasing difficulty, with the exception of reading comprehension questions (found in the Verbal Reasoning and English domains), which are arranged according to the order in which the subject matter appears in the text. The questions in all three domains are designed to test abilities that are theoretically required for success in academic studies.

The Verbal Reasoning sections test verbal abilities including vocabulary, logical thought processes (syllogistic reasoning), the ability to analyze and understand complex reading comprehension passages, the ability to think clearly and methodically, and the ability to formulate a thought and express it in writing in a manner that is well-reasoned and precise. This domain comprises three sections: the writing task (one section) and two sections, each made up of 23 multiple-choice questions (for a total of one essay and 46 multiple-choice questions) consisting of the following scales: Verbal analogies, sentence completions, syllogistic questions, and reading comprehension (based on either a single item of a very short passage or on a short text).

The Quantitative Reasoning sections test the ability to use numbers and mathematical concepts for solving quantitative problems, as well as the ability to analyze data presented in a variety of formats, such as tables or graphs. This domain comprises 2-3 sections (depending on the test version), each made up of 20 multiple-choice questions (for a total of 40 questions). The current version comprised of three sections – questions and problems, table comprehension and graph comprehension.

The English sections test proficiency in the English language, including vocabulary and reading comprehension on an academic level. This domain comprises two sections, each made up of 22 multiple-choice questions (for a total of 44 questions). Overall, the PET comprises 130 multiple-choice questions and one writing task, and in Study 1, the factor analysis was performed on 12 different test item types in 12 scales (see below).

Ensuring Test Item Quality: Before the questions in any given section can be included in an operational test, their quality, fairness, and ability to differentiate

between examinees with greater and lesser abilities must be established. Therefore, each test contains two pilot sections, which are used for quality control purposes. Questions from those sections that meet statistical and other criteria are likely to be included in sections that will be incorporated into future tests and scored; other questions, which do not meet these criteria, will not be included in future tests. All sections that are scored are made up of questions that have undergone this quality check. In order to maintain the integrity of this process, examinees are not informed which sections are scored and which are not. It should be noted here that our factor analysis used only the six operational sections of the selected test version.

Data Analysis

The 12 scales of the PET were subjected to exploratory factor analysis. The EFA was conducted using Maximum Likelihood Estimation (MLE), where the squared multiple coefficient (SMC) was used as a prior for the communality estimates. Finally, a Promax (oblique) rotation was done. This procedure has an advantage over techniques such as Principal Component Analysis (PCA) because it controls for measurement error to identify latent factors underlying the manifest variables, rather than simply condensing information provided by those variables. The rotation method used allows the factors to correlate with one another, and therefore improves the model and its interpretability (Lawley & Maxwell, 1971). The number of extracted factors was identified using both Kaiser's (1960) criterion (Eigenvalue>1) and Cattel's (1966) Scree Plot method (extraction of factors above an inflection point on a graph of plotted Eigenvalues).

Results

Descriptive Statistics

Table 1 presents the means and standard deviations of the variables (item types). The right-hand side of Table 1 presents Cronbach's alpha (Cronbach, 1951) reliabilities for each indication. The reliability of the writing task was estimated at 0.50, based on estimates of examinees who were tested in the US (Breland, Kubota, Nickerson, Trapani & Walker, 2004).

Table 1 – Means, Standard Deviations and Reliabilities for Test Scores

Variable	Mean	STD	Number of Items	Reliability
Questions and problems	19.75	6.31	32	0.86
Table comp.*	2.68	1.10	4	0.42
Graph comp.*	2.39	1.31	4	0.60
Writing task	14.26	3.86	24	0.50
Verbal analogies	8.51	2.47	12	0.68
Sentence completion	3.88	1.46	6	0.55
Logic	6.15	1.93	10	0.56
Reading comp.* – single item	3.46	1.40	6	0.46
Reading comp.*	7.02	2.99	12	0.76
Sentence completion (English)	10.86	4.08	16	0.86
Restatements	5.37	2.32	8	0.79
Reading comp.* (English)	13.06	4.46	20	0.84

*=Comprehension

Exploratory Factor Analysis Results

Table 2 presents the Pearson's intercorrelation matrix among the various variables representing the test sub-domains.

Table 2 – Intercorrelations among the 12 Test Variables

	QP	TC	GC	WT	VA	SC	LG	RC-s	RC	SC-e	RE	RC-E
QP	1.00											
TC	0.48	1.00										
GC	0.55	0.38	1.00									
WT	0.46	0.29	0.31	1.00								
VA	0.57	0.34	0.37	0.51	1.00							
SC	0.46	0.30	0.33	0.43	0.50	1.00						
LG	0.55	0.36	0.42	0.43	0.50	0.46	1.00					
RC-S	0.43	0.29	0.35	0.40	0.43	0.40	0.48	1.00				
RC	0.52	0.35	0.41	0.49	0.50	0.44	0.47	0.44	1.00			
SC-E	0.50	0.32	0.36	0.40	0.47	0.48	0.41	0.40	0.48	1.00		
RE	0.53	0.33	0.36	0.41	0.48	0.43	0.43	0.43	0.49	0.79	1.00	
RC-E	0.56	0.36	0.42	0.44	0.47	0.44	0.46	0.45	0.56	0.76	0.76	1.00

QP=Questions and problems, TC=Table comprehension, GC= Graph comprehension, WT=Writing task, VA=verbal analogies, SC= Sentence completion, LG=Logic, RC-S=Reading comprehension-single item, RC=Reading comprehension, SC-E=Sentence completion – English, RE=Restatements, RC-E=Reading comprehension – English.

The correlations presented in Table 2 have not been corrected for reliability. As will be demonstrated below, reliability has an indirect influence on the model through errors of measurement related to the different variables, and hence, factor analysis was conducted on this intercorrelational matrix without correcting it for reliability.

The eigenvalues suggested a two-factor solution for the PET scales. The first eigenvalue was 12.96 and the second eigenvalue was 1.69, yielding a first-to-second eigenvalue ratio of 7.67. Scale loadings after Promax (oblique) rotation for the two-factor solution are presented in Table 3; values greater than 0.4 are considered significant.

Table 3 – Scale Loadings on Factors in the Two-Factor Solution

Variable	Factor 1	Factor 2
Questions and problems	.71*	.10
Table comp.*	.50*	.04
Graph comp.*	.57*	.04
Writing task	.57*	.08
Verbal analogies	.65*	.09
Sentence completion	.57*	.08
Logic	.70*	.01
Reading comp.* – single item	.52*	.13
Reading comp.*	.56*	.19
Sentence completion (English)	-.04	.94*
Restatements	.08	.81*
Reading comp.* (English)	.15	.78*

*=Comprehension

The two factors emerging from Table 3 may be interpreted as a *General Reasoning Skill* factor and an *English Proficiency* factor. The *General Reasoning Skill* factor comprised the writing task, reading comprehension, verbal analogies, sentence completion, logic, math problem-solving, and table and graph comprehension; in other words, the Quantitative Reasoning and Verbal Reasoning domains. The *English Proficiency* factor comprised all of the English section scales: sentence completion, restatement, and reading comprehension.

However, an examination of the scree plot reveals that the number of data-points above the "break" was three, suggesting it is necessary to consider a three-factor solution as well. The three-factor model is also consistent with the theoretical structure of the test. Therefore, we imposed a three-factor solution on the data. Scale loadings after Promax (oblique) rotation for the three-factor solution are presented in Table 4.

Table 4 – Scale Loadings on Factors in the Three-Factor Solution

Variable	Factor 1	Factor 2	Factor 3
Questions and problems	.20	.06	.63*
Table comp.*	.03	.03	.54*
Graph comp.*	.00	.02	.66*
Writing task	.68*	.04	-.04
Verbal analogies	.65*	.05	.07
Sentence Completion	.61*	.05	.02
Logic	.51*	-.01	.25
Reading comp.* – single item	.49*	.11	.07
Reading comp.*	.46*	.17	.15
Sentence Completion (English)	.00	.93*	-.03
Restatements	.08	.80*	.02
Reading comp.* (English)	.07	.77*	.10

*=Comprehension

The three factors emerging from Table 4 may be interpreted as a *Verbal Reasoning* factor, a *Quantitative Reasoning* factor, and an *English Proficiency* factor. The *Verbal Reasoning* factor comprises the writing task, verbal analogies, sentence completion, and reading comprehension (either a single question on a very short passage or several questions on a longer passage). The *Quantitative Reasoning* factor comprises math problem-solving, and table and graph comprehension, while the *English Proficiency* factor comprises English sentence completion, restatement, and reading comprehension. In other words, when imposing a three-factor solution, the *General Reasoning Skill* factor, found previously, is further divided into two factors, representing the test's scholastic domains. The scale loading values indicate that the three-factor solution is also highly probable.

Study 2

Overview

As both a two-factor solution and a three-factor solution may be justified by the results of the EFA, these solutions were further tested as competing models in the confirmatory factor analysis conducted in Study 2. Specifically, we tested three competing models that were all based on theoretical and empirical considerations. The first was a one-factor model which, based on previous studies that found high *g* loadings in scholastic aptitude tests, assumes that all test domains actually measure a single construct (a *g*-like factor). The second, based on the two-factor solution of the EFA, was a two-factor model that consists of Verbal and Quantitative Reasoning as representative of an "ability" factor and English as representative of a "knowledge" factor. Finally, the third model was the three-factor model based on the PET's three content domains, which may be expected to act as three separate factors, as reflected in the scree plot in Study 1.

Using CFA also allowed us to apply additional statistical procedures to examine the psychometric properties of the test. Among those procedures was an analysis of correlations between test domains, an assessment of reliability (indicator and complex, which will be discussed in more detail below), an assessment of validity (convergent and divergent), and an estimation of the extracted variance – the amount of variance that is captured by an underlying factor in relation to the amount of variance due to measurement error. Therefore, more than reevaluating the structure of the PET, this study examined how changes in the format in general, and the addition of the writing task in particular, influenced the internal structure of those capabilities that are assessed by the test and additional psychometric properties. In the second stage, we used several parameters offered by the CFA to revise the internal structure model and ultimately, proposed a modified and improved hierarchical, non-standard model, which more accurately matched the data and pointed to additional relations between the sub-domains measured by the test.

Method

Materials

Test Version Used in the Present Study: In order to conduct factor analysis, we selected a different PET administration. The test version that was chosen was

administered in September 2012 (in Hebrew). The test was given to 3,704 examinees. Males comprised 38% of the sample and females 62%; average age was 21.8 years (standard deviation of 2.39, median 21.8). Native Israelis made up 90% of the sample. About 2.8% of examinees reported that their socioeconomic status was very much higher than average, 28% much higher than average, 41.8% slightly above average, 16.5% slightly below average, 7.6% much below average, and 2.8% very much below average. Statistical comparisons indicated that this sample was representative of the population of examinees during the September test administration over the previous 10 years. This sample's characteristics are also highly similar to those of the sample in Study 1, and it may be assumed that both samples represent the general examinee population. The Quantitative Reasoning section of this version of the test included only two scales – graph comprehension and math problem-solving. Thus, in this study, the CFA was conducted on 11 scale scores.

Results

Descriptive Statistics

Table 5 presents the means and standard deviations of the variables (item types). The right-hand side of Table 5 presents Cronbach's alpha reliabilities for each scale.

Table 5 – Means, Standard Deviations and Reliabilities for Test Scores (Study 2)

Variable	M	STD	Number of Items	Reliability
Questions and problems	20.70	5.73	32	0.84
Graph comp.*	4.68	1.84	8	0.55
Writing task	14.35	4.08	24	0.50
Verbal analogies	7.81	2.28	12	0.60
Sentence completion	3.60	1.44	6	0.46
Logic	6.86	1.84	10	0.55
Reading comp.* – single item	4.21	1.43	6	0.50
Reading comp.*	7.41	2.91	12	0.77
Sentence completion (English)	12.06	3.66	16	0.85
Restatements	5.66	1.90	8	0.67
Reading comp.* (English)	13.29	4.22	20	0.82

*=Comprehension

Correlations among the Various Sections of the Test

Table 6 presents the Pearson's intercorrelation matrix among the various variables representing the sub-domains of the test.

Table 6 – Intercorrelations among 11 Test Variables (Study 2)

	QP	GC	WT	VA	SC	LG	RC-s	RC	SC-e	RE	RC-E
QP	1.00										
GC	0.57	1.00									
WT	0.40	0.30	1.00								
VA	0.54	0.41	0.46	1.00							
SC	0.39	0.31	0.36	0.41	1.00						
LG	0.50	0.42	0.40	0.47	0.43	1.00					
RC-S	0.43	0.38	0.40	0.44	0.42	0.52	1.00				
RC	0.50	0.41	0.47	0.49	0.43	0.46	0.48	1.00			
SC-E	0.46	0.36	0.39	0.46	0.30	0.40	0.41	0.48	1.00		
RE	0.50	0.38	0.38	0.46	0.38	0.43	0.44	0.50	0.74	1.00	
RC-E	0.52	0.42	0.42	0.49	0.38	0.46	0.46	0.58	0.81	0.75	1.00

QP=Questions and problems, GC=Graph comprehension, WT=Writing task, VA=verbal analogies, SC= Sentence completion, LG=Logic, RC-S=Reading comprehension – single item, RC=Reading comprehension, SC-E=Sentence completion – English, RE=Restatements, RC-E=Reading comprehension – English.

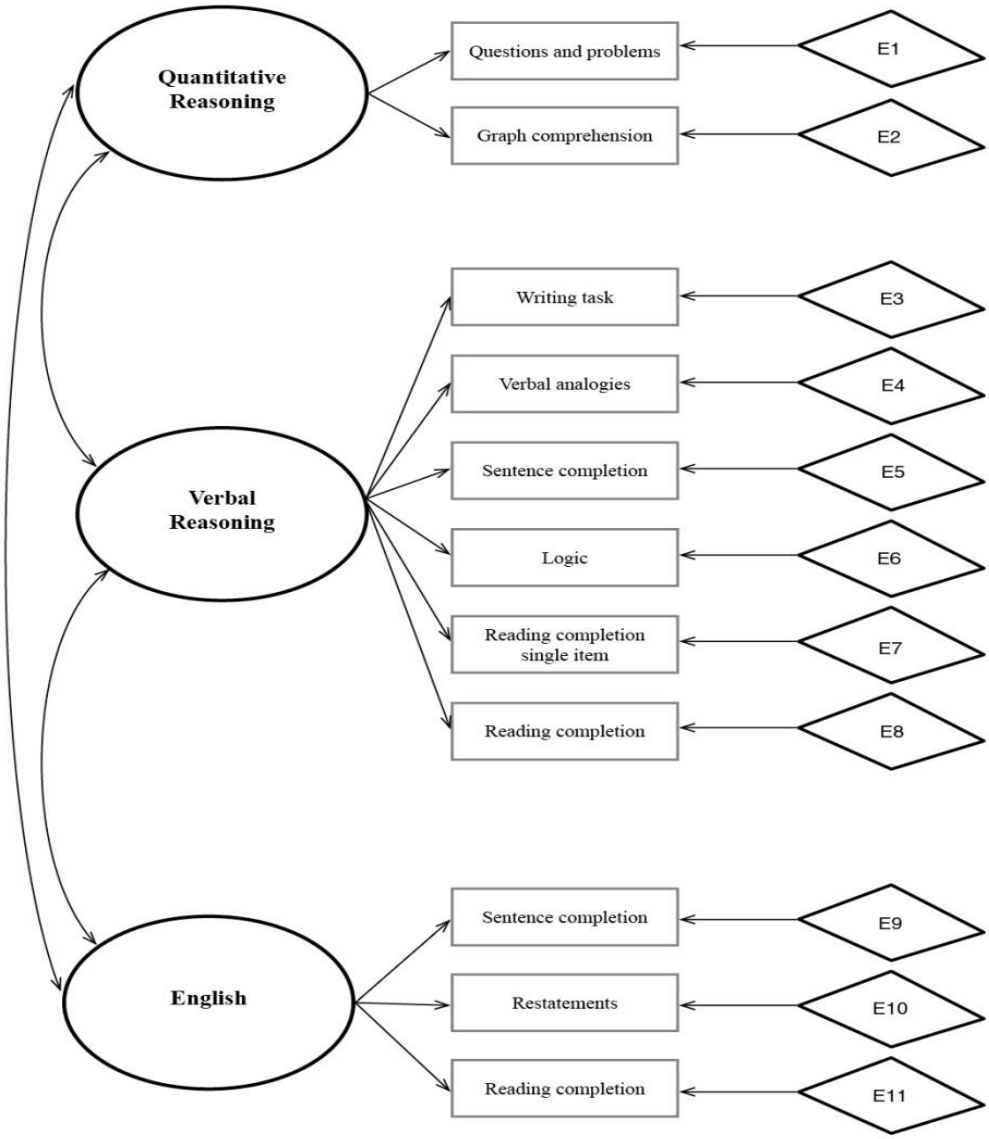
The correlations presented in Table 6 have not been corrected for reliability. Intercorrelations among the three test domains and the general PET score ranged from moderate to high. The correlation between Quantitative Reasoning and Verbal Reasoning domains was .67, and .55 with the English domain. The correlation between the Verbal Reasoning and English domains was .65. The total PET score correlated .89 with the Verbal Reasoning domain, .83 with the Quantitative Reasoning domain, and .87 with the English domain.

The Internal Structure of the Test Scales – Testing Competing Models

We hypothesized that the best fit indices would emerge for the three-factor model, which consists of three latent factors that match the test domains, where the 11 scales are manifest indicators (i.e., test item types, each of which is explained by the domain to which it belongs). Figure 1 is a schematic presentation of this hypothesized content-derived model. In accordance with convention, Figure 1 presents the manifest variables (scales) inside rectangles, latent factors inside ellipses, and errors of measurement inside rhombuses.

In total, there are 11 manifest endogenous variables, i.e., the 11 scales that are measured by the test; three latent exogenous factors, which are the three test domains; and 11 exogenous variables of errors of measurement that belong to each measured indicator. As the result of the scale indeterminacy problem, the variance of each latent factor was fixed to 1, leaving 25 model parameters that need to be estimated (the arrows): 11 loading coefficients of the factors' scales, 11 variances of error, and three correlations between factors. We compared this model to the competing one- and two-factor models.

Figure 1 – Schematic Presentation of the Three-Factor Hypothesized Model



Overall Goodness of Fit Indexes. When using Confirmatory Factor Analysis, it is accepted practice to apply several statistical tests in order to assess goodness of fit between the model and the data (Shur, 2006). A report on goodness of fit indexes frequently includes several of the following measures (Kline, 2010; Hatcher, 1994). The *Chi-Square Test* indicates the difference between observed and expected covariance matrices; values closer to zero indicate a better fit and a smaller gap between expected and observed covariance matrices. One difficulty with the chi-square test of model fit, however, is that it is overly sensitive to large sample sizes (Bollen, 1989), such as that in our study (Type II error, Gatignon, 2010). As a result, several other measures of fit were also examined. The *Root Mean Square Error of Approximation (RMSEA)* avoids issues of sample size by analyzing the discrepancy between the hypothesized model, with optimally chosen parameter estimates, and the population covariance matrix (Hooper, Coughlan & Mullen, 2008). This measure ranges from 0 to 1, with smaller values indicating better model fit. A value of 0.08 or less is indicative of acceptable model fit and a value of 0.06 or less is considered excellent fit (Hu & Bentler, 1999). In addition, the *Goodness of Fit Index (GFI)* is a measure of fit between the hypothesized model and the observed covariance matrix. The measure ranges from 0 to 1, with a cutoff value of 0.90 generally indicating acceptable model fit. The *Normed Fit Index (NFI)* and the *Non-Normed Fit Index (NNFI)* were also considered. The NFI analyzes the discrepancy between the chi-square value of the hypothesized model and the chi-square value of the null model (Bentler & Bonnett, 1980). However, NFI tends to be negatively biased (Bearden, Sharma & Teel, 1982). The NNFI resolves some issues of negative bias, though NNFI values may sometimes fall beyond the 0 to 1 range (Bentler 1990). The values for both the NFI and NNFI should range between 0 and 1, with a cutoff of 0.95 or greater indicating a good model fit (Hu & Bentler, 1999). Finally, the *Comparative Fit Index (CFI)* analyzes the model fit by examining the discrepancy between the data and the hypothesized model, while adjusting for issues of sample size inherent in the chi-square test of model fit and the NFI. Values range from 0 to 1, with larger values indicating better fit; a CFI value of 0.90 or larger is generally considered to indicate acceptable model fit and a value of 0.95 or higher is considered to indicate excellent fit.

Goodness of Fit Indexes for the Competing Models. Table 7 presents goodness of fit indexes that were received for each of the three competing models.

Table 7 – Goodness of Fit Indexes for the Various Competing Models

Model	χ2	*Df*	RMSEA	GFI	NFI	NFFI	CFI
3-Factor Model	502	41	0.0551	0.9763	0.9753	0.9695	0.9773
2-Factor Model	776	43	0.0678	0.9630	0.9619	0.9538	0.9639
1-Factor Model	3,139	44	0.1378	0.8212	0.8458	0.8095	0.8476

The CFA yielded the highest fit indexes for the hypothesized three-factor model, indicating significantly better fit than the two other models. The one-factor model shows unacceptable fit indexes in all parameters. The two-factor model shows acceptable to good fit indexes; however, they were still inferior to those of the hypothesized three-factor model.

In particular, except for the chi-square index (which was significant), all the other measures show a very good fit with the three-factor model. Again, when the sample is large (as in the current study), the chi-square test will very frequently be significant, even if the model provides a good fit (James, Mulaik, & Brett, 1982). In real-world situations, therefore, it has become common practice to seek a model with a relatively small chi-square value, rather than one with a non-significant value.

Just how small the chi-square must be depends on the degrees of freedom (*df*) associated with the analysis. If the model analyzed fit perfectly with the data, then chi-square has an expected value equal to the *df* (Marsh, Balla & MacDonald, 1988). Therefore, in most cases, a requirement for significance is exchanged for a requirement for a relationship between the chi-square and its degrees of freedom, so that a value of less than 3 for the chi-square/*df* ratio indicates a good fit. It should be noted, however, that because the value is also heavily affected by sample size (Marsh et al., 1988), one should be doubly cautious when using this measure. For our three-factor model, this measure's value was 12.24, which as mentioned above, does not show a good fit. Therefore, in the following analyses, we attempted to further improve the three-factor model and to arrive at a better representation of the internal structure of the test, taking into consideration both the empirical finding and theoretical considerations.

Scale Loadings on the Factors. Table 8 presents the loadings, including the *t* values for the model based on item type. Non-standardized values are presented first, while the last column presents standardized values.

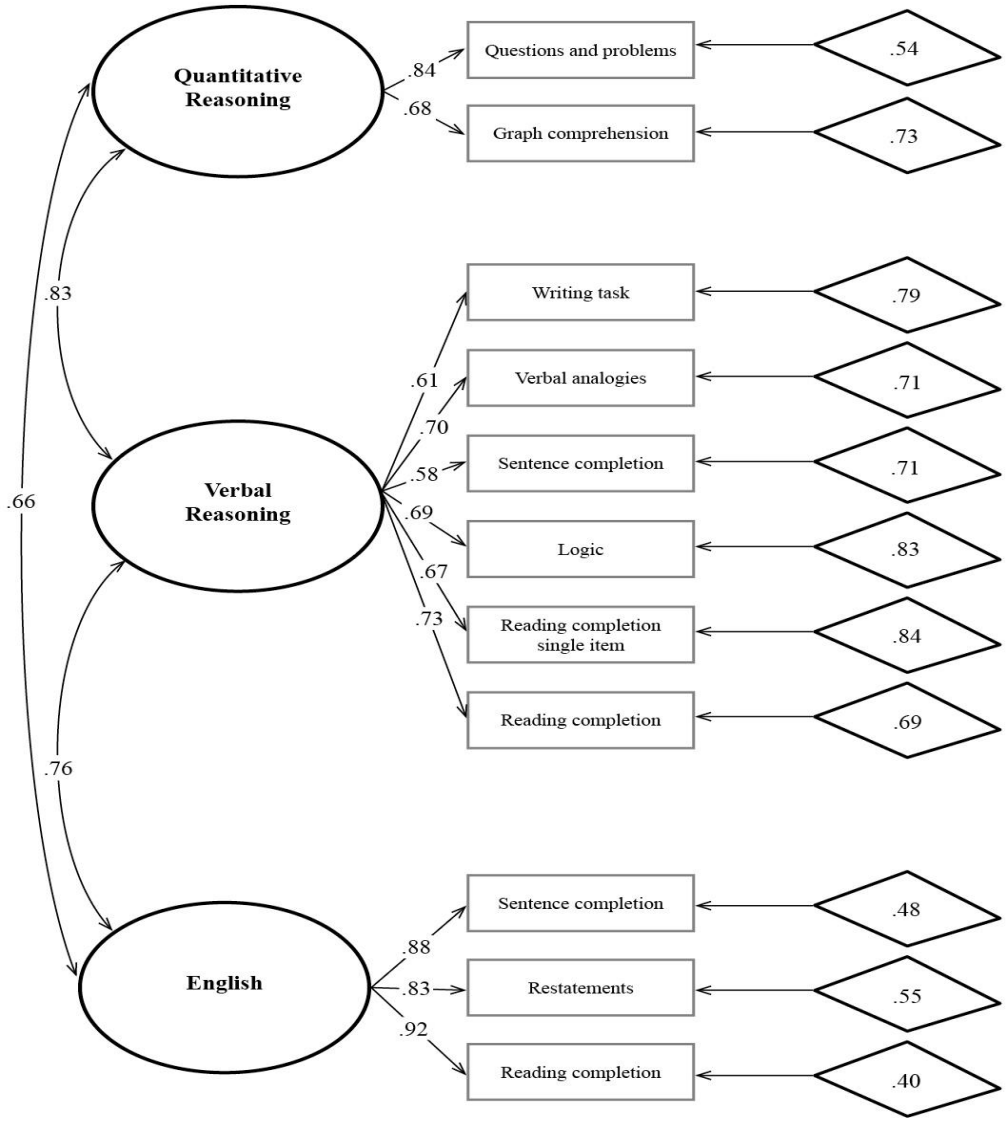
Table 8 – Scale Loadings in the Three-Factor Model

Variable	**Non-Standardized Values**			**Standardized Loading**
	Loading	Standard Error	*t* value*	
Questions and problems	4.83	0.090	53.67	0.84
Graph Comp.**	1.25	0.029	42.56	0.68
Writing task	2.48	0.064	38.63	0.61
Verbal analogies	1.61	0.035	46.58	0.70
Sentence Completion	0.84	0.023	36.74	0.58
Logic	1.27	0.028	45.07	0.69
Reading comp.** – single item	0.96	0.022	43.56	0.67
Reading comp.**	2.12	0.044	48.61	0.73
Sentence Completion (English)	3.21	0.049	65.89	0.88
Restatements	1.58	0.026	60.86	0.83
Reading comp. (English)	3.87	0.055	70.68	0.92

*All values are significant at the 0.001 level, **=Comprehension, ***=English

Figure 2 presents the results of Table 5 (the last column of standardized values) in graph form. The figure also includes the variance values of the errors and correlations between the factors (which will be discussed below).

Figure 2 – Confirmatory Factor Analysis Results for the Three-Factor Model



Correlations among the Factors. As explained above, in addition to loading coefficients (from the factors to the indicators and error variances), the model provides estimates of correlations among factors, shown in Figure 2. The correlations among factors received from the CFA are high relative to the correlations based on examinees' raw scores. This inconsistency seemingly raises a doubt as to whether the three test domains were actually found.

The answer to this question lies in one of the characteristics of the confirmatory model. The confirmatory model allows an error factor that influences each of the indicators. This error factor embodies a reliability correction, and therefore, in actuality, the correlations presented in Figure 2 are comparable to the raw correlations

after correcting for reliability. In order to test this claim empirically, the reliability of each factor in the confirmatory analysis was calculated and the resemblance between the correlations was examined after correcting for reliability. The reliabilities were .85, .85, and .92 for the Quantitative Reasoning, Verbal Reasoning, and English domains, respectively. Correcting the correlations to the reliability of the variables, we receive: $r(verbal, quantitative) = 0.79$, $r(quantitative, English) = 0.62$, and $r(verbal, English) = 0.74$. After this correction, the values received become very similar to those presented in Figure 2.

Reliability and Validity

One of the most important advantages offered by latent-variable analysis is the opportunity to assess the reliability and validity of the study variables. The following section presents the estimated reliability and validity of indicators and factors as they arose from the Confirmatory Factor Analysis. Combined, these procedures provide evidence concerning the extent to which the indicators used in the study are producing reliable data and are measuring what they are intended to measure.

Reliability

Indicator Reliability: The reliability of an indicator variable is defined as the square of the correlation between the latent variable and that indicator. In other words, the reliability indicates the percentage of variation in the indicator that is explained by the factor that it is supposed to measure (Long, 1983). This value is equal to the square of the standardized loading of the factor's variable.

Composite Reliability: Similarly, when performing Confirmatory Factor Analysis, it is possible to calculate a composite reliability index for each latent variable included in the model. This index is analogous to the Cronbach's alpha, and reflects the internal consistency of the indicators measuring a given factor. Calculating composite index reliability (Fornell & Larcker, 1981) is done as follows:

$$Composite\ Reliability = \frac{(\sum L_i)^2}{(\sum L_i)^2 + \sum Var\,(E_i)}$$

where L_i = the standardized factor loading for that factor and

$Var(E_i) = 1 - L_i^2$ = the error variance associated with the individual indicator variables.

Estimate of Variance Extracted: Fornell and Larcker (1981) discussed an index called the variance extracted estimate, which assesses the amount of variance that is explained by an underlying factor in relation to the amount of variance due to measurement error. The formula they suggested was:

$$Variance\ Extracted = \frac{\sum L_i^2}{\sum L_i^2 + \sum Var(E_i)} = \frac{\sum L_i^2}{n}$$

with the same indications as in the previous formula (where *n* is the number of variables). Table 9 presents these measures:

Table 9 – Indicator Reliability, Domain Reliability and Extracted Variance

Domain	Variable	Indicator Reliability*	Domain Reliability	Extracted Variance
Quantitative Reasoning	Questions and problems	0.71	.74	.58
	Graph Comp.**	0.46		
Verbal Reasoning	Writing task	0.37	.83	.44
	Verbal analogies	0.50		
	Sentence Completion	0.34		
	Logic	0.47		
	Reading comp.** – single item	0.45		
	Reading comp.**	0.53		
English	Sentence Completion (English)	0.77	.91	.77
	Restatements	0.69		
	Reading comp.** (English)	0.84		

* Calculated as the square of the standardized loading of the factor, **=Comprehension

The column showing the reliability scale in Table 9 should be comparable to the reliability column (last column) in Table 5. A comparison of the two tables shows that, with the exception of *English reading comprehension*, the reliabilities calculated in the CFA model were lower than those that were calculated directly using the Cronbach's alpha formula. It should be noted that in Table 5, the reliabilities were calculated using the Cronbach's alpha formula on the basis of individual items that served as the indicator – information that was included in the factor analysis.

As opposed to indicator reliability, the domain reliability presented in Table 9 is very similar to the Cronbach's alpha reliability of the three factors, which was calculated once again on the basis of individual items that make up the indicator. (One can see that the reliability of the Quantitative Reasoning domain in Table 9, which is made up of just two scales, is relatively less accurate than the other two domains, which comprise a larger number of scales.)

Fornell and Larcker (1981) suggested that it is desirable that constructs exhibit estimates of 0.50 or larger, because estimates less than 0.50 indicate that variance due to measurement error is larger than the variance captured by the factor. Our results show that the Verbal Reasoning and English domains meet this threshold but that the Quantitative Reasoning domain does not.

Validity

Convergent validity and discriminant validity are usually associated with an analysis conducted using the Multitrait-Multimethod (MTMM) approach, in which multiple constructs are assessed using more than one assessment method (Campbell & Fiske, 1959). It has been claimed that the MTMM approach provides a stronger test of convergent (and discriminant) validity than is afforded by the procedure here (for example, Widaman, 1985; Schmitt & Stults, 1986; Netemeyer, Johnston & Burton, 1990). Nonetheless, the procedure used here sheds some light on the internal structure of the test and is useful in cases in which it is not possible to use the MTMM approach.

Convergent Validity. Convergent validity is demonstrated when different instruments are used to measure the same construct, and scores from these different instruments are strongly correlated. In the present study, convergent validity was assessed by reviewing the *t* test for factor loadings. If all the factor loadings for the indicators measuring the same construct are statistically significant (greater than twice their standard error), this is viewed as evidence supporting the convergent validity of those indicators (Anderson & Gerbing, 1988). As seen in Table 5, all of the loadings were found to be statistically significant, supporting the convergent validity of the proposed construct.

Discriminant Validity. Discriminant validity is demonstrated when different instruments are used to measure different constructs, and the correlations between the

measures of the different constructs are relatively weak. A test displays discriminant validity when it does not measure a construct that it was not designed to measure.

In the current study, a chi-square difference test was carried out to assess the discriminant validity of the three proposed factors. We compared the model we received to each of three other models (different from the competing models we tested earlier), which differ from the original model in one way: in each of them, two of the factors are forced to be fully correlated (r=1). Discriminant validity is demonstrated in the event that the chi-square value of the original model is significantly lower than that found in any of the other models, thus suggesting that a model in which the two constructs are assumed to be distinct (but correlated) is preferable to a model in which the factors are equal (Anderson & Gerbing, 1988; Bagozzi & Phillips, 1982). For purposes of checking discriminant validity, Table 10 presents the data for the four models' chi-square measures.

Table 10 – Test of Discriminant Validity

Model	$\chi^2(df)$	$\Delta\chi^2(\Delta df)$
Original model	502(41)	
$r(quantitative, verbal) = 1$	775(42)	273(1)
$r(English, quantitative) = 1$	1248(42)	273(1)
$r(English, verbal) = 1$	1248(42)	1838(1)

A value of 10.8 is needed to determine significance at a level of 0.001 for one degree of freedom. All of the differences in Table 10 are (much) larger than this value and thus, support the discriminant validity of the different test domains.

The Revised Model

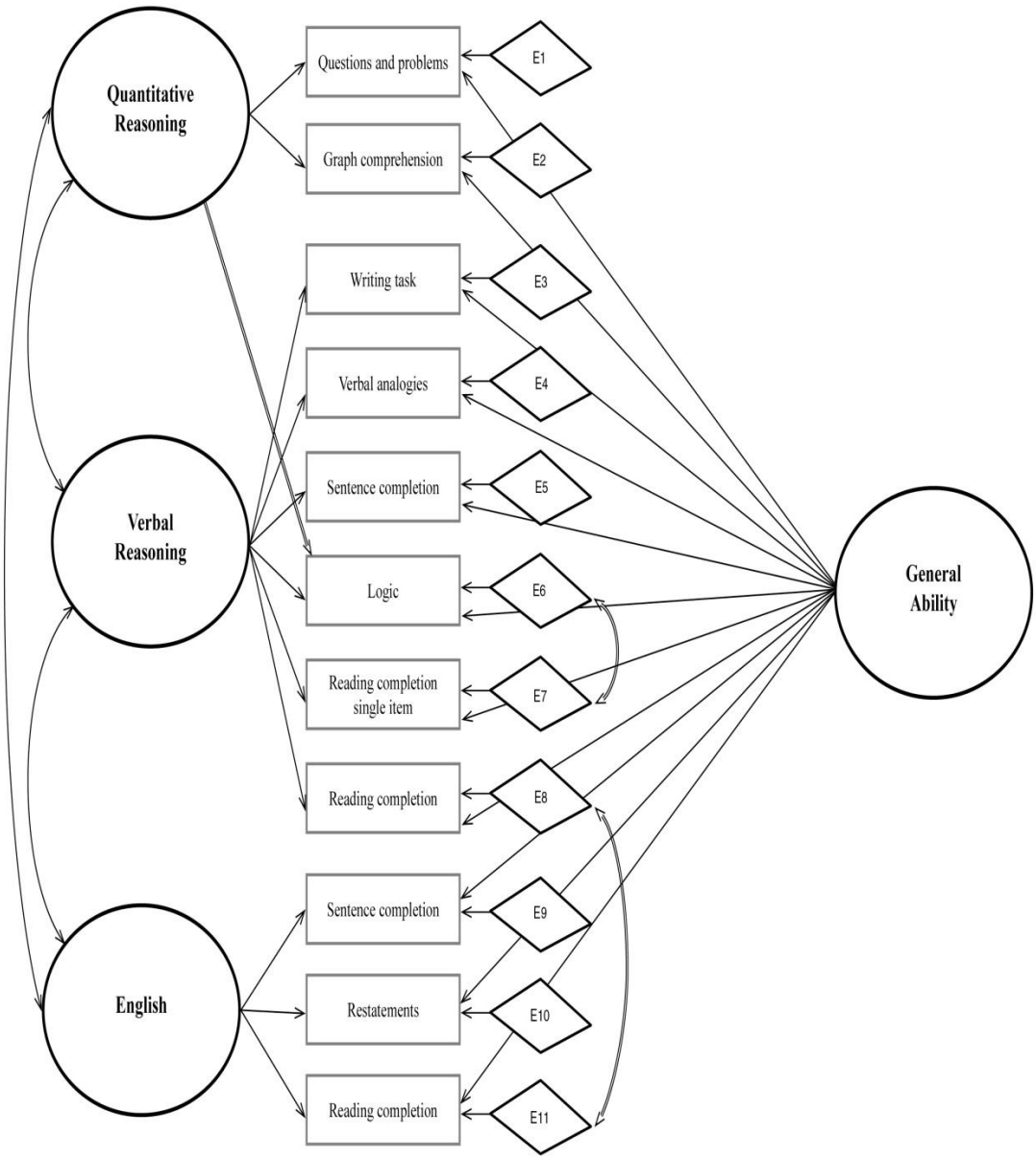
Based on the analyses described above, we propose an improved model that is a variation on the three-subject content model outlined above. We were encouraged to formulate and test a revised model by several CFA parameters that indicated shortcomings of the original model: (a) indications that a scale may be influenced by more than one factor (multidimensional indicator), (b) indications of covariance between errors, and (c) high correlations among the three factors.

To address the first two shortcomings, we relied on modification indexes – the *Wald Test* and the *Lagrange Multiplier Test*. The Wald Test helps to identify paths and

links that could potentially detract from the model. The Lagrange Multiplier Test identifies paths and links between variables that might need to be added to the model. Using these indexes, we propose a model in which the *logic* scale, which was originally attributed to the *Verbal Reasoning* domain, is in fact influenced by two domains: *Verbal Reasoning* and *Quantitative Reasoning*. In addition, we revised the model to allow for covariance between scales: (a) between *reading comprehension* and *reading comprehension in English*, and (b) between *logic* and *reading comprehension-single item*. In the revised model, this is done by allowing the corresponding errors of these variables to correlate among themselves.

To address the third shortcoming, we reconstructed the revised model to be a hierarchical two-layer model, with a *General Ability* factor affecting all 11 scales. This was done in the face of high correlations among the factors that hinted at the existence of a latent general ability factor. The revised model is presented in Figure 3 (differences from the original three-subject content model are indicated by thick arrows).

Figure 3 – Schematic Presentation of the Revised Non-Standard Model



Goodness of Fit Indexes. Table 11 presents goodness of fit indexes for the revised model (third column). For purposes of comparison, the second column presents the values received for the original three-subject content model.

Table 11 – Goodness of Fit Indexes for the Original and Revised Models

	Original Model	Revised Model
Chi-Square	502 (DF=41), Pv<0.0001	181 (DF=30), Pv<0.0001
Chi-Square/DF	12.24	6.03
RMSEA	0.0551	0.0368
GFI	0.9763	0.9911
NFI/NNFI	0.9753/0.9695	0.9911/0.9864
CFI	0.9773	0.9926

This table shows that the revised model is an improvement over the original three-subject content model – all of the measures show a better fit. It should be noted that the chi-square value is still statistically significant, but the quotient between the chi-square value and degree of freedom is much smaller (however, still above the threshold of 3).

Scale Loadings on the Factors. Table 12 presents loading values, including their significant values for the revised model. The table presents standardized loading values on the content factor (left-hand side) and loading values for the *g* factor (right-hand side).

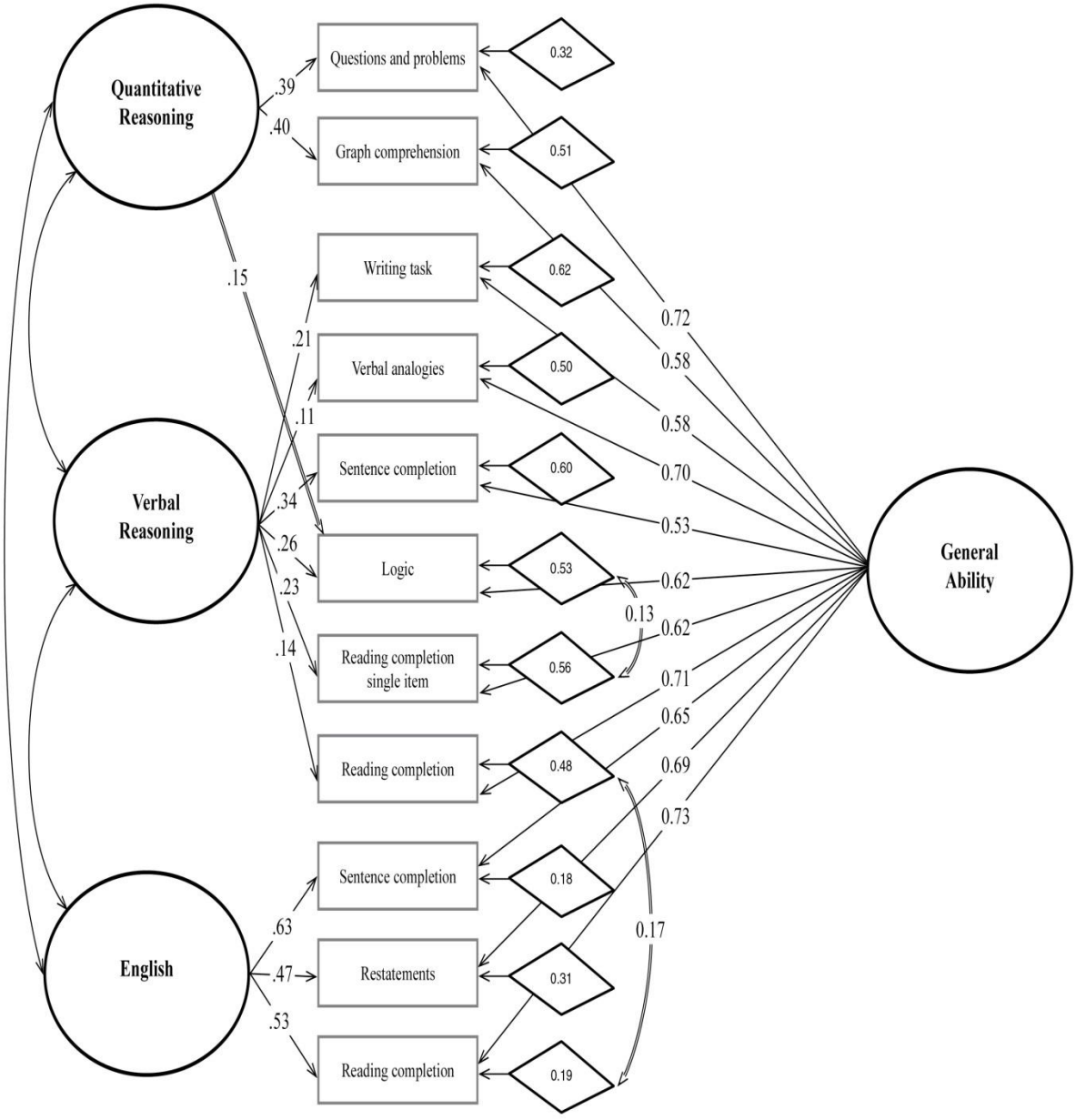
Table 12 – Scale Loadings in the Revised Model

Domain	Scale	Standardized Loading	Loading on g
Quantitative Reasoning	Questions and problems	0.39	0.72
	Graph comp.*	0.40	0.58
	Logic	0.15	0.62
Verbal Reasoning	Writing task	0.21	0.58
	Verbal analogies	0.11	0.70
	Sentence completion	0.34	0.53
	Logic	0.25	0.62
	Reading comp.* – single item	0.23	0.62
	Reading comp.*	0.14	0.71
English	Sentence completion (English)	0.63	0.65
	Restatements	0.47	0.69
	Reading comp.* (English)	0.53	0.73

*Comprehension

The results presented in Table 12 are presented schematically in Figure 4. The figure also notes the variance values of the errors and the correlations among the factors and errors.

Figure 4 – Confirmatory Factor Analysis Results for the Revised Non-Standard Model



When comparing Figure 4 and Figure 2, one can see that overall, the loadings of the scales on the content factors become much smaller. Also, the loading for the *logic* scale of the *Verbal Reasoning* factor, which was .69 in the original three-subject content model, is now divided in the revised model: .15 for *Quantitative Reasoning* and .25 for *Verbal Reasoning*.

Discussion

The present study used an Exploratory Factor Analysis (Study 1), followed by a Confirmatory Factor Analysis (Study 2), to gain insight into the construct validity of the Psychometric Entrance Test (PET) for higher education in Israel, as represented by the internal factorial structure of the scholastic aptitudes it measures. In addition, it examines various aspects of the test's convergent and discriminant validity (Bentler, 1990). Our analyses showed that the model that best fits the internal structure of the abilities represented by the PET was a two-layer hierarchical model, in which a *General Ability* (*g*) factor affects all measured scholastic abilities; however, the three specific content domains – *Verbal Reasoning*, *Quantitative Reasoning*, and *English* – still add unique explanatory value over and above general ability.

In general, the findings support the construct validity of the PET and identify three empirical factors, corresponding to the test's three existing content domains. Specifically, the CFA supported the internal three-factor construct, where each scale could only be loaded on the factor to which it was theoretically attributed. For this model, referred to as the original three-subject content model, four of the five estimation measures gave values in the acceptable to good range. That is, the original three-subject content model was fairly close to the description of the factor construct of the various test scales. At the same time, when some of the test scales were allowed to load on more than one factor or to correlate with factors other than those to which they were theoretically attributed (expressed in the model as a correlation between errors of measurement), the goodness of fit indexes improved. Also, adding a general ability factor to the model together with, but not instead of, the three content factors improved its explanatory power. This model, referred to as the revised model, yielded excellent fit indexes. It should be noted that, despite the high loadings of all the scales on the *General Ability* factor, a single-factor model alone could not accurately describe the internal structure of the test. Combining a general ability factor and the three content factors in the model yielded the best-fitting model.

These findings are consistent with those from other studies conducted in the US. Rock et al., (1988), who looked at high-stakes tests with similar constructs, concluded that they were a measure of more than general ability alone, but also of the individual scholastic aptitudes they were designed to assess. In highlighting the necessity of the three specific content factors for accurately describing the construct

measured by the PET, our findings also support the conclusion reached by Coyle and Pillow (2008), who claimed that while the SAT and ACT were highly *g* loaded, both tests predicted GPA from non-*g* factors. Thus, with regard to the main question of whether the test presents a single measure of overall ability or a measure of the fields of knowledge that are being tested (Koening, Frey & Detterman, 2008), our findings indicated the former is probably not the case, and that a combination of the two theoretical conceptualizations probably results in the best-fitting model.

With regard to loadings of specific scales on the different factors, as expected, the *writing task* was loaded on the *Verbal Reasoning* factor, a finding that generally supports the way in which the score in the Verbal Reasoning domain is currently calculated. Moreover, there was a significant correlation between *Hebrew reading comprehension* and *English reading comprehension*. This finding could suggest the existence of general linguistic factors that influence functioning in both the mother tongue and a foreign language. Indeed, previous studies point to the transfer of proficiencies from the mother tongue to a foreign language, and their results indicate that aptitude in the mother tongue is one of the best predictors of aptitude in a foreign language (Kahn-Horwitz, Shimron & Sparks, 2005; Sparks et al., 1997; Sparks, Patton, Ganschow & Humbach, 2009; Sparks, Patton, Ganschow, Humbach & Javorsky, 2008).

Finally, it was found that the *logic* scale was also loaded to a certain degree on the *Quantitative Reasoning* factor. It could be that this finding, which was not among the initial assumptions, points to a network of linkages between language functioning and quantitative reasoning performance. Specifically, it is possible that syllogistic reasoning, which is seen in formal logic items, also depends to some extent on the skills needed to solve problems in the Quantitative Reasoning domain, i.e., various analytical and abstract thinking abilities. Consistent with this explanation, the literature offers two differential paths of explaining cognitive processes that are based on deductive reasoning: the syntactic path, which is linguistic-labial based, and the visual-spatial path, which requires spatial manipulative ability, a cognitive skill that is closely linked to mathematical and quantitative capability. Studies show that these paths are actually two separate systems that are linked at the level of brain function (Goel, Buchel, Frith & Dolan, 2000).

Implications of the Study and Directions for Future Research Limitations

Before discussing the implications of these findings, one must acknowledge the limitations of the present study. First, the study looked at two specific versions of the PET, which were administered on certain dates. Although the samples were large and representative, perhaps the findings should be replicated with other samples that use different test versions and are administered at different times of the year. Replicating the findings under these circumstances would provide significant support for the validity of the proposed model. In addition, the studies are based on data collected from Israeli samples, using an Israeli test. Israeli candidates for higher education may differ from their counterparts in other countries in several ways (e.g., older age due to military service, degree of religiosity, etc.). However, although this may limit the cross-cultural generalizability of the findings, it should be noted that the PET is highly equivalent to other worldwide tests used for the same purpose, and that the Israeli population is generally similar to that in other Western countries. Nevertheless, replicating the findings in additional international samples is needed to support their generalizability. Finally, the measures of convergent and discriminant validity are based solely on factor analysis, and although they support the test's validity, they are only an initial and partial examination.

Implications and Conclusions

Our findings indicate that scholastic aptitude tests for higher education probably do not measure a single general ability factor, but rather that the scholastic domains assessed by the test have unique explanatory power. This finding has theoretical implications for the conceptualization of such tests.

Because the findings of the current study support the expected theoretical construct of the PET and its theoretical division into three different content domains, they reinforce the rationale for the test and the way in which the various scores are calculated. Specifically, it appears that, as the test developers intended, the PET indeed assesses linguistic and quantitative skills as well as English proficiency. In light of these findings, it also seems that the way in which the scores are calculated today, which allows examinees and the educational institutions to receive a score with an emphasis on verbal reasoning, a score with an emphasis on quantitative reasoning, and a general score, matches the empirical findings and offers a valid tool for accepting or rejecting applicants on a differential basis to various fields of study.

Furthermore, the findings support the test's face validity, and confirm that the changes made to the test over the years did not harm the theoretical constructs on which the test is based, and that the current format can continue to be used.

Summary and Directions for Further Research

Our study focused mainly on construct validity of the PET. Future studies will be needed to examine the predictive validity of the test's new format, and to determine whether there were changes in predictive validity as a result of changes in its structure in general, and the role of the writing task – which has very different attributes than the other test components – in particular.

In addition, the study should be repeated using additional samples in order to assess the stability of our conclusions regarding the latent structure underlying scholastic aptitudes. This may include testing the cross-cultural generalizability of the findings with other similar tests and with samples from other countries.

Moreover, the current study used Confirmatory Factor Analysis to assess convergent and discriminant validity. Future research could use other widely accepted tests of validity such as MTMM, in order to provide further support for the convergent and discriminant validity of the factors identified in the present study.

In summary, the present study, which tested various criteria of validity for the PET, and specifically for its factorial construct, offers good initial support for the test's validity and makes it reasonable to assume that the changes made to the test's content and construct over the years did not impede its evaluative validity. Future studies will look at the question of whether the changes actually improved the test's predictive validity and quality.

References

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives*, *10*, 18.

Anderson, J.C. & Gerbing, D.W. (1988). Structural equation modeling in practice: A review *and* recommended two-step approach. *Psychological Bulletin*, *103*, 411-423.

Bagozzi, R.P. & Phillips, L.W. (1982). Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly*, *27*, 459-489.

Bearden, W.O., Sharma, S., & Teel, J.E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research*, *19*, 425-430.

Beller, M., (1990). Tree versus geometric representation of tests and Items. *Applied Psychological Measurement*, *14*(1), 13-28.

Beller, M. (2001). Admission to higher education in Israel and the role of the Psychometric Entrance Test: Educational and political dilemmas. *Assessment in Education: Principles, Policy & Practice*, *8*(3), 315-337.

Beyth-Marom, R., Beller, R. Brown, H., Brennan, R. Gafni, N., Hambleton, R., Cohen, Y., Nevo, B. & Tzelgov, Y. (1998). The role of the Psychometric Entrance Test in higher education candidate selection. (RR No. 248.) Jerusalem: National Institute for Testing & Evaluation (in Hebrew).

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-46.

Bentler, P.M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin*, *112*(3), 400-404.

Bentler, P.M., & Bonnett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303-316.

Brelnad, H. Kubota, M., Nickerson, K., Trapani, C., & Walker, M. (2004). New SAT writing prompt study: Analysis of group impact and reliability. (College Board Research Report No. 2004-1 ETS RR-04-03.) New York: College Entrance Examination Board.

Budesco, D. (1985). Factor analysis of a set of university entrance tests. (RR No. 21.) Jerusalem: National Institute for Testing & Evaluation (in Hebrew).

Cahalan-Laitusis, C., Cook, L. L., & Aicher, C. (2004, April). Examining test items for students with disabilities by testing accommodation. Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cattell R, B. (1966).The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276.

Cook, L., Eignor, D., Steinberg, J., Sawaki, Y., & Cline, F. (2014). Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on a reading comprehension assessment. *Journal of Applied Testing Technology*, *10*(2), 1-33.

Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing g. *Intelligence*, *36*(6), 719-729.

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

Fornell, C. & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18,* 39-50.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science*, *15*(6), 373-378.

Gatignon, H. (2010). *Statistical Analysis of Management Data* (Chapter 4), 2nd Edition, New York, NY: Springer.

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage*, *12*(5), 504-514.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika, 33*(4), 469-506.

Hatcher, L. (1994). *A Step by Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling*. Cary, N.C.: SAS Institute Inc.

Hooper, D., Coughlan, J. & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Journal of Business Research Methods, 6*, 53–60.

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Huynh, H., & Barton, K. E. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education*, *19*(1), 21.

James, L.R., Mulaik, S.A. & Brett, J.M. (1982). *Causal Analysis*. Beverly Hills: Sage.

Kahn-Horwitz, J., Shimron, J., & Sparks, R. L. (2005). Predicting foreign language reading achievement in elementary school students. *Reading and Writing*, *18*, 527–558.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.

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. (RR No. 264.) Jerusalem: National Institute for Testing & Evaluation.

Kennet-Cohen, T., Oren, C., Turvall, E., & Cohen, Y. (2013). Predictive validity of the candidate selection system in Israeli universities' law schools. (RR No. 285.) Jerusalem: National Institute for Testing & Evaluation (in Hebrew).

Kleper, D., Allaluf, A., Turvall, E., Oren, C., & Pronton, M. (2015). Two-layer quality control of PET scores by demographic characteristics. (RR No. 415.) Jerusalem: National Institute for Testing & Evaluation (in Hebrew).

Kleper, D. & Turvall, E., (2016). A meta-analysis of the predictive validity of the Psychometric Entrance Test. (RR 16-02.) Jerusalem: National Institute for Testing & Evaluation (in Hebrew).

Kleper, D., Turvall, E., & Oren, C. (2014). Predictive validity of the PET in predicting higher first year GPA. (RR 403.) Jerusalem: National Institute for Testing & Evaluation (in Hebrew).

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling* (3rd Ed.). New York, New York: Guilford Press.

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence, 36*(2), 153-160.

Lawley, D. N. (1971). Factor analysis as statistical method (No. 519.5 L3 1971).

Long, J.S. (1983). Confirmatory factor analysis: A preface to LISREL. Sage University Paper Series on Quantitative Application in the Social Sciences, 07-033. Beverly Hills, CA: Sage.

Marsh, H.W., Balla, J.R. & MaDonald, R.P. (1998). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

Netemeyer, R.G., Johnston, M.W. & Burton, S. (1990). Analysis of role conflict and role ambiguity in a structural equations framework. *Journal of Applied Psychology, 75*, 148-157.

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Oren, C., Kennet-Cohen, T., Turvall, E., & Allalouf, A. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema*, *26*(1), 117-126.

Rock, D. A., Bennett, R. E., & Jirele, T. (1988). Factor structure of the graduate record examinations general test in handicapped and nonhandicapped groups. *Journal of Applied Psychology, 73*(3), 383.

Sattath, S. & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*(3), 319-345.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schmitt, N. & Stults, D.M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1-22.

Shur, D.D. (2006). Exploratory or confirmatory factor analysis? *Statistics and Data Analysis, 200-31*. In the Proceedings of the 31st Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc.

Sparks, R. L., Artzer, M., Ganschow, L., Patton, J., Siebenhar, D., & Plageman, M. (1997). Prediction of Foreign Language Proficiency. *Journal of Educational Psychology, 89*(3), 549-561.

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2008). Early First-language reading and spelling skills predict later second-language reading and spelling skills. *Journal of Educational Psychology, 100*(1), 162–174.

Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Applied Psycholinguistics, 30*, 725–755.

Steinberg, J., Cline, F., & Sawaki, Y. (2011). Examining the factor structure of a state standards-based science assessment for students with learning disabilities. (College Board Research Report ETS RR-11-38.) New York: College Entrance Examination Board.

Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.

Zwick, R. (2007). *College Admission Testing*. Arlington, VA: National Association for College Admission Counseling.