דוח המרכז

387

Using Public Opinion to Inform the Validation of Test Scores

Tzur M. Karelitz

March 2013

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

דוח מרכז 387

ISBN:978-965-502-176-9

Using Public Opinion to Inform the Validation of Test Scores

Tzur M. Karelitz

National Institute for Testing and Evaluation (NITE), Israel

The author would like to thank Baruch Nevo, Yoav Cohen, Naomi Gafni, Carmel Oren, Jerry Levinson, Avi Allalouf, Anat Ben-Simon, Edo Litmanovitch and the NITE staff for their help with this research.

Contents

Executive summary

Standardized tests, such as the Psychometric Entrance Exam (PET) are commonly the subject of a heated public debate. Public opinion can affect decision regarding the test's content and structure, its administration and scoring, and the interpretation and usage of test scores. Regardless of whether a test is psychometrically valid or not, public pressure can influence policy-makers to consider the test invalid for its intended purpose.

This paper discusses why and how test developers should consider what people think about the test. We argue that it is worth exploring the disparities between the intended purposes of the test and what people think the test actually measures - namely, the "face validity" of the test. Although face validity is considered a subjective judgment which is not indicative of the test scores' predictive or construct validities, we believe that it might offer a unique contribution to the development of a sound interpretive argument. The disparity between the intentions of the test developers and people's perceptions of the test can lead to various misinterpretations and misuses of test scores. Test developers must not only explain how test scores should be interpreted and used (Kane, 2006), but also why certain alternative interpretations and uses are inappropriate and may lead to undesirable consequences.

To support our claims, we present findings from a study about perceptions of the PET. Following the methodology outlined by Nevo (1985), we surveyed future, present and past PET examinees, faculty and admissions officers from institutes of higher education, and other public figures who deal with educational issues. We also surveyed the PET developers and other measurement experts. Apart from background information, the data we collected included multiple-choice and open-ended questions about: (a) the relevance of the PET to its intended use, (b) the extent to which the PET measures abilities important for success in higher education, (c) the extent to which people agree with various claims for and against the PET, (d) ways that the PET scores are used that differ from their intended use, and (e) preferences among tools and methods for selecting students to higher education. The paper discusses the study findings, their implications,

and how they were used to support proposed changes in the structure, content, administration and score interpretation of the PET.

Overall the results show the complex nature of opinions that the public holds about the test. In general, the public acknowledge the quality of the test design, administration and scoring. However, most people have negative views about the test's content and time constraints. Moreover, a large proportion of the public does not think the test is needed or that it successfully fulfills its intended purpose. These views vary somewhat between future students, current and past students, and academic staff. The results also show that some aspects of the test are particularly problematic in the public eyes. We show that some of these problems can be significantly reduced by future changes to the test design and administration.

Introduction

Companies routinely ask for customer feedback — whether they produce laptops, movies, or provide banking services — public opinion matters. Surveys, polls and focus groups can help reveal perceptions about the product, identify areas for improvement and reduce future problems. However, the application of these methodologies for test improvement is limited, especially in education (Stricker, Wilder & Bridgeman, 2006). This is unfortunate given that standardized tests are the subject of a heated public debate (e.g., Phelps, 2003). Public opinion can affect the test's content and structure, its administration and scoring, the interpretation and usage of test scores, and the very existence of the test. Regardless of whether a test is psychometrically valid or not, public pressure can influence policy-makers to consider the test invalid for its intended purpose.

This paper discusses why test developers should consider what people think about the test. We argue that it is worth exploring the disparities between the intended purposes of the test and what people think the test actually measures - namely, the *Face Validity* (FV) of the test. We discuss two ways that traditional validity analysis can benefit from the study of FV, as well as other benefits related to the public relations of the test. To illustrate our approach, we provide examples from a large study about the Psychometric Entrance Test (PET) to institutions of higher education in Israel.

The paper is organized as follows. In the Introduction we discuss the concept of FV and situate it within the state-of-the-art of validity research. This part of the introduction outlines the theoretical reasoning for our study. Next, we provide necessary background information about higher education in Israel. In the Methods section we describe the study we conducted. In the Results and Discussion section we present the main findings and discuss how they illustrate the usefulness of public opinion for test score validation. We conclude with final remarks and recommendations.

Background on Face Validity

Face validity represents the way a test is perceived by the examinees, the users of the test and the general public. A common interpretation is that a test has good FV if, taken at face value, it appears to measure what it is intended to measure. This perception

of the test lacks empirical credence as it is based on the subjective evaluation of laypeople and not on a statistical model. Consequently, researchers consider FV as a weak form of validity evidence, as Downing and Haladyna (2004) note, "the appearance of validity is not validity". The main criticism against FV is that it is not an objective measure of validity (Mosier, 1947; Downing, 2006). Following Mosier's call to banish the term to "outer darkness", the discussion about validity has shifted to other issues and FV has all but disappeared (it is absent from major publications such as the Standards for Educational and Psychological Testing). Interestingly, Moiser's point may have been somewhat misunderstood. He did not argue that researchers should avoid studying FV, but rather objected to the vagueness of the term. His final recommendation was: "Anyone intending to use the term should, instead, describe fully the concept which he originally intended to denote by *face validity*" (1947, p. 205).

Although a test's FV does not indicate the existence of predictive validity or construct validity, it is of great importance in and of itself (Anastasi, 1988; Chronbach, 1970). For example, Nunnally (1967) suggests that FV can be seen as an indirect approach to the measurement of content validity. Turner (1979) argues that some tests must be face valid for other tests to be construct valid, otherwise a validity analysis will be logically flawed. According to Nevo (1985), FV is important because it can affect (a) examinees' motivation to prepare and perform well, (b) the willingness of potential examinees to take the test, (c) the level of dissatisfaction of examinees with low scores, (d) the opinions of decision makers regarding the use of the test, and (e) the opinions of the general public, the media and the judicial system. Efforts to strengthen FV are usually aimed at increasing the acceptance of the test among the examinees and other stakeholders (Kane, 2006).

The FV of a test is reflected by the popularity of the test, which is affected by the test's public relations, by the "movers and shakers" of public opinion, and by the influence of various interest groups. When FV is low, the popularity of the test might diminish to the point where its continued existence could be at stake. Low FV creates a unique threat that test developers must consider when studying the validity of the test. Test developers can do a lot to ensure that the test maintains desirable psychometric properties. This is crucial at a professional level and helps avoid many threats to validity

in general. However, the public is usually unaware of, or uninterested in, the psychometric properties of the test. For the layperson, forming an opinion about the test is motivated more by satisfaction (or lack thereof) with their test outcomes and less by the test's reliability coefficient. Consequently, it is much easier, and more common, for the public to criticize the test using the type of arguments that are the focus of FV than to criticize its psychometric properties. Test developers must realize that the opinion of the public matters, as it has the power to determine the fate of the test. For example, if examinees or users have a choice between multiple tests, they are likely to choose the test that has a higher FV. If there is only one test, then examinees can protest or take other political action to advocate the development of a different test or of different criteria for decision making; they can even take steps to abolish the test completely. The bottom line is that low FV threatens the existence of the test itself, and that this threat has more to do with the test's public relations than with its psychometric rigor.

The rationale for the current study is that FV should be studied routinely throughout the life cycle of a test. We believe that studying the FV of a test adds useful information for establishing the validity of the test and can be helpful for test developers in their efforts to improve the public's opinion of the test. We do not propose that the study of a test's FV is sufficient for validation purposes. Validity analysis is always warranted, and can only be augmented by the analysis of FV. We explain below how this perspective on FV can be integrated into the current framework of validity theory.

A "Facelift" for Face Validity

There has been little discussion about FV since Moiser (1947). Consequently, some advances in validity theory have not been applied to the concept of FV. For example, Moiser spoke of FV in terms of the test content, whereas current theories appropriately identify that it is the usage of test scores that we wish to validate. In this paper we attempt to inspect the concept of FV in terms of current validity theories.

The analysis of test validity focuses on *validity as an argument* (Cronbach, 1988; Kane, 2006). According to this approach, test validators should construct an interpretive argument by specifying the network of assumptions and inferences that underlie the proposed interpretations and uses of test scores. That is, one should explicate the logical argument that leads from observed performance to conclusions about examinees and to

any decisions based on these conclusions. The validity argument is used to evaluate the interpretive argument by studying its coherence and by using evidence to support or refute its underlying assumptions and inferences. The evaluation of evidence focuses on the test content, its internal structure, the underlying response processes, the test's outcomes and their connections to other relevant variables. In order to rigorously evaluate the plausibility of the interpretive argument, one must also consider the plausibility of alternative interpretations and uses of test scores.

We see two reasons why FV can be useful for the current validity framework. First, to evaluate the plausibility of the proposed interpretive argument, test validators need to juxtapose their arguments against various alternatives. A good source for alternatives can be the beliefs held by examinees, test users and decision makers regarding the test scores' interpretations and uses. In that sense, studying the FV of a test can be seen as a process of collecting and evaluating evidence about alternative inferential networks. Second, as indicated above, low FV creates a unique threat to the existence of the test. If the public rejects the use of the test, we can no longer claim that the test is valid. Test validators should collect evidence to evaluate the extent to which public opinion pose a threat to the validity of test scores. Therefore, it makes sense to apply the interpretive argument framework specifically to the study of the public's perceptions about the test. We describe below a systematic approach for constructing and evaluating an interpretive argument about the perceived validity of a test. We report a study that uses this approach to evaluate the FV of a large-scale standardized test used for the admissions process to institutes of higher education (IHE) in Israel.

An Operational Definition of Face Validity

To construct the interpretative argument we begin by extending the operational definition of FV given by Nevo (1985) in the following mapping sentence:

A RATER who is a(n) [testee/ nonprofessional user/ interested individual (public)] RATES a(n) [test item/ test/ battery] BY EMPLOYING a(n) [absolute/ relative] TECHNIQUE AS [very suitable (or relevant)… unsuitable (or irrelevant)] FOR ITS INTENDED USE.

The definition focuses on the match between qualities measured by the test and the role of the test scores in the decision making process. This captures the essence of validity— the appropriateness of the interpretation and usage of test scores. The definition indicates that the relevant evidence stems from the subjective judgments of individuals about various components of the test. Although this mapping sentence was suggested more than 25 years ago, it has not been widely applied. Next, we describe an extension of the first two facets of this definition.

We extend the scope of the first facet to include groups we believe are relevant to the study of FV. Specifically, if the test has qualities that contribute to its validity, then people who are more familiar with the test and its purpose will evaluate its FV more favorably than those who are less familiar with it. To study whether this is true in reality, researchers need to focus on different groups. For example, the term "testee", can be expanded to describe three types of examinees: past, present and future. Past examinees have already participated at an earlier administration of the test, present examinees are those who have just completed the real test for the first time, and future examinees are those who have not yet participated in a real exam. Note that once an examinee participates in the test he or she is considered past examinee even if they decide to re-take the test. Some of the past examinees may have already received the results of their test, and possibly know how their test scores have been used. For example, if the test is an academic admission test, some of the past examinees may have already matriculated into their chosen institutions. The hypothesis behind the proposed change is that the three groups represent different levels of familiarity with the test and its purpose. For example, a high school student who just registered to take the SAT knows less about the test than an applicant who has been studying to the test for the past 3 months and who has just completed their first real exam. Both of them know less about academic studies than a past examinee who is currently a student at an IHE (see also, Secolsky, 1987).

Other groups of interest include various levels of test users. For example, for the SAT, the user groups might be (1) admissions staff, (2) academic administrators, and (3) professors. Again, these groups represent various levels of interaction with the test scores, and perhaps even different conceptions of the role of the test for academic admissions. Another important group includes politicians, media people, public figures,

and others who help shape and inform (or misinform) public opinion. Finally, we believe it is important to include the test developers and related professionals in the sample. Traditionally, the way professionals view the test is considered evidence of content validity. We think this information can also be used in conjunction with the FV judgments of examinees, users and laypeople. It is obvious that a test would have distinctively better FV in the eyes of its developers, but (a) it is worth verifying this hypothesis with evidence, and (b) the developers perceptions can be used as an upper "anchor case" to compare to other groups.

The second facet focuses on the structure of the test. For diagnostic purposes, it is important to study the FV of various components of the test. For example, one could ask respondents to connect the purpose of the test to every type of item on the test, or to the various sub-tests. This information could be useful later in focusing efforts to improve the test or its public relations. The remaining facets could stay as they are, which gives us the following operational definition:

A RATER who is a(n) [interested individual, past/present/future examinee, professional/nonprofessional user, public figure, professional test developer] RATES a(n) [item type, sub-test, test, battery] BY EMPLOYING a(n) [absolute/ relative] TECHNIQUE AS [very suitable (or relevant)… unsuitable (or irrelevant)] FOR ITS INTENDED USE.

The definition suggests that there could be a variety of perceptions about the FV of the test, ranging from completely negative to completely positive. It is also possible that this range of perceptions will correlate well with attitudes regarding the continued usage of the test. It is reasonable that those who perceive the test to be invalid for a particular purpose might also advocate the discontinuation of the test.

An Interpretive Argument for Face Validity of Higher Education Admission Tests

Building on the operational definition of FV given above and on the framework presented by Kane (2006), we propose an interpretive argument for the FV of an admission test to IHE. That is, we present the network of assumptions and inferences needed to reach conclusions about the face validity of the test. In order to form an

educated opinion about the test, a preliminary condition is that individuals are familiar with:

a. the purpose of the test (i.e., why the admissions process requires a selection test).
b. how test scores are used in the admissions process (e.g., how test scores are combined with other information to reach a decision based on some criteria).
c. the test's content, structure, mode of administration and scoring process.

The various pieces of knowledge about the test are strongly inter-related. They are shaped by experiences with the test, and they can change with time. Obviously, not everyone who forms an opinion about the test possesses all these pieces of knowledge, and not all the knowledge they do possess is accurate. For example, someone who erroneously thinks that a HE admissions test measures 8th grade math might question the relevance of the test scores for admissions. Alternatively, someone who knows the test has a section on verbal abilities might think the test is biased against immigrant populations and thus question its use for HE admissions.

Moreover, possessing this knowledge does not mean that one thinks positively about the test, since you can know what the purpose of the test is and at the same time think that the test does not meet that purpose. A more detailed chain of inference is needed to construct an interpretive argument about the test's FV. The components of the argument we provide in Figure 1 identify the assumptions and inferences underlying the use of test scores for admissions decisions based on Kane's (2006) framework. The four inferences we wish to make lead from (a) observed performance to observed score, (b) observed score to universe score (i.e., true score), (c) universe score to conclusions about ability level, and (d) conclusions about ability level to an admissions decision. Within each inference in Figure 1 we have listed the relevant underlying assumptions (see also Lyrén, 2009). Unlike other conceptions of validity, the assumptions concern the public's perceptions about the test, rather than the actual properties of the test.

Figure 1. The interpretive argument about the FV of an admission test

Scoring: from observed performance to observed score

1. people think that the test scores are calculated accurately and consistently, and appropriately represent performance on the test.

Generalization: from observed score to true score

1. people think that the test scores are based on a representative sample from the universe of observations (i.e., different testing conditions, test forms or raters).
2. people think that the number of items is large enough to provide a reliable estimate of the true score.

Extrapolation: from true score to ability level

1. people think that the test items require abilities that are relevant for successful academic performance.
2. people do not think there are other abilities that are relevant for successful academic performance but are not measured by the test.
3. people think that there are no ability-irrelevant sources of variability that would bias the interpretation of scores as measures of ability level.

Decision: from conclusion about the ability level to decision about admission

1. people think that the conclusions about ability levels, as measured by the test, are meaningful for admission decisions.
2. people think that students with low abilities, as measured by the test, are likely to perform worse academically than students with high abilities.
3. people think that using the conclusions about ability levels for admissions decisions does not have a negative social impact.
4. people do not think that conclusions about ability levels are used inappropriately for other purposes.

The validity argument aims to evaluate the appropriateness of these inferences and assumptions. This requires the collection of relevant evidence to support or refute the claims in Figure 1. Obviously, some of the terms and nuances are foreign to the general public (e.g., true score, reliability) and therefore they cannot be asked directly in this form. Instead, one can derive statements and questions that reflect these assumptions and use those to survey people's knowledge, attitudes and beliefs.

More specifically, the assumption about scoring inferences would probably be false if people have negative perceptions about the expertise and integrity of the organizations involved in developing, administrating and scoring the test. People's knowledge (or lack thereof) about the scoring procedures (calibration, parallel forms equating, doglegging, etc.) might also shape their beliefs about how well the scores reflect performance.

The assumptions that underlie inferences about generalization from observed score to true scores depend on people's knowledge and beliefs about variations in testing conditions. For example, a belief that the test tends to be easier on certain administration dates might lead people to discredit the generalization of test scores. In addition, the length of the test can be perceived in various ways. Some might think a long test provides better measurement because of the multiplicity of measurements, some might think a long test provides worse measurement because examinees get tired, while others are unaware of the relation between test length and quality of measurement.

The assumptions that underline inferences about extrapolation from true score to ability level depend on people's perception of the factors that influence performance on the test. People might question the relevance of the measured abilities for the purpose of predicting performance in academic environment. They might think the abilities are relevant but not important, or they might think the abilities are important but not measured well (or not measured at all) by test items. Alternatively, people might think that there are relevant abilities that are not measured by the test, although they should be. In addition, people might think that performance is more influenced by other factors such as the ability to work under time-pressure, the ability to use test-wiseness, the motivation to perform well on the test, the socio-cultural background of examinees and so on. The test coaching industry might influence these perceptions by claiming to provide "tricks" that can help one score higher on the test. The differential performance of various groups might cause people to think the test is biased or unfair and therefore discredit inferences about ability level.

The assumptions that underline inferences about admissions decisions based on conclusions about ability level depend on people's perception of the appropriateness of the test as criterion for admission. Because admission decisions have a great impact on

people's life, and consequently, on society at large, people's perceptions of using the test for admission are influenced by their own performance or by the performance of those who are related to them. Moreover, people might base these perceptions on single cases ("I know someone who did very well on the test but did very poorly as an undergraduate"). In making such claims people ignore the complex network of factors that influence academic performance. Finally, people might encounter situations in which the admissions test scores are used for other purposes, which were not originally intended by the test developers. Such cases might cause people to have negative perceptions of the test, not because it is inappropriate for admissions decisions, but because it is inappropriately used for other purposes.

The process of FV analysis includes collecting evidence to evaluate the extent to which people prescribe to this chain of inferences and assumptions. This is helpful for (a) evaluating the extent to which the test has FV, (b) identifying alternative score interpretations and uses, and (c) designing solutions for the aspects in the chain of inferences that seem most problematic. In order to facilitate interpretation, the collected evidence should include information about (a) the background characteristics of the respondents, (b) the factors that influence their perceptions about the test, and (c) the personal and social consequences of having such perceptions.

From such evidence, test developers can not only learn about the public perception of the test, but also learn how negative perceptions could be addressed. Studying FV systematically is particularly important for large-scale, high-stakes, standardized tests that are usually under public and political scrutiny. Such studies are mostly relevant for tests that have been used for some time and about which the public has already established opinions. Test developers could use such information to explore possible modifications to the test or to focus on how to improve their public relations efforts. The study presented in this paper illustrates how the FV argument can be evaluated empirically. The next section provides the context in which we conducted our study.

Admission to Higher Education in Israel

In 1948, when Israel was founded, two universities served a population of about 800,000 people. Today, 7 national Universities, one Open University and almost 60

colleges, serve a population of 7.8 million. Overall, the proportion of the population (age 25-64) who possess an academic degree is among the highest in the world (29%, similar to the USA). In the past 20 years, the proportion of students among the population in the relevant age group (20-24) has increased from 23% to about 47%[1]. The number of undergraduate students has almost tripled in the past two decades, a growth which was facilitated by the spurt of academic colleges in Israel. Two decades ago, 85% of all students studied in the universities, and less than 10% studied in academic colleges. Today, only 34% of the students study in one of the universities, and 47% attend academic colleges or teacher preparation institutions (another 19% study at the Open University). The tuition for the universities and academic colleges is subsidized by the government and averages around $2,500 a year. The tuition in private colleges can be three times that amount.

The admissions process to most IHE in Israel is competitive, and only about two thirds of the 75 thousand applicants matriculate each year. In most institutions, admissions is based on two criteria— the high school graduation exams and the Psychometric Entrance Exam (PET). The high school graduation exams, the Bagrut, are a set of assessments conducted during the final years of high school. Students are tested on at least 7 different topics, including Math, English, Hebrew, History, Literature, Bible studies and Civil studies. Most IHE (with the exception of the Open University), require a completed Bagrut certificate for admissions, and have additional criteria regarding the required level of math and English. Many students choose to do additional exams, or take advanced exams in some topics, to receive additional credits that count towards admissions to IHE. Each year, about 50% of the cohort pass all the requirements and receive the Bagrut certificate. Students who do not complete the Bagrut in high school can obtain the certificate later (an additional 3%). The Bagrut exams are not standardized in the sense that testing conditions are not constant across administrations, scores are not calibrated across cohorts or topics, and often there are problems with test security.

[1] In comparison, in 2006 about 25% of Canadian citizens of the relevant age group were students [http://www4.hrsdc.gc.ca/.3ndic.1t.4r@-eng.jsp?iid=56#M_1]. The data regarding Israel's higher education system, is drawn from the Israel Council For Higher Education, "Facts and Numbers 2011", 2012 [http://www.che.org.il/download/files/factsandnumbers.pdf] (in Hebrew)

The PET is developed, administered and scored by the National Institute for Testing and Evaluation (NITE) [2]. This non-profit organization was founded by the universities of Israel in the early 1980's and has been operating on a financially independent basis ever since. The PET's purpose is to provide a standardized admissions test for higher education focusing on academic abilities (for more information on the subject see Attali, & Goldschmidt, 1999; Beller, 1994, 2001). The PET consists of six operational sections and two pilot sections. The pilot sections are used for developing new items and for equating purposes, and therefore examinees are not scored on these sections (see Allalouf, 1999 for information about scoring and equating). The six operational sections consist of two sections in each of three areas – verbal reasoning, quantitative reasoning, and English.

Every year, NITE develops about 8 paper and pencil versions of the PET in Hebrew. Each form contains about 200 multiple-choice items. NITE translates 3 forms for the benefit of test takers in other languages (Arabic, Russian, English, French and Spanish, overall about 40% of the examinees in any given year). NITE also administers adapted tests for examinees with special needs (including computerized adaptive tests). There are 5 administration dates each year, and people can take the test as many times as they wish (although they cannot complete two consecutive tests). The cost of the test for each individual examinee is around $130. NITE also provides an extensive preparation handbook and publishes one solved PET form after each administration.

The testing conditions are standardized across locations and years and test security is held at a very high standard. The PET's standardized scores range between 200 and 800, with a mean around 535 (SD≈100). The score is a weighted average of the three domains (40% verbal, 40% quantitative and 20% English). The PET has established desirable psychometric properties (Kennet-Cohen, Bronner, & Oren, 1995). It has high reliability coefficients (Cronbach alpha around .95 for the whole test) and predictive validity coefficients (correlation with first year GPA= .46, compared with .38 for the Bagrut). In terms of effect size, the validity of the admission score (a simple average of the Bagrut and the PET scores) is equal to a standardized difference of 0.8 standard deviations, which is considered a strong effect (Oren, Kennet-Cohen & Bronner, 2007).

[2] NITE internet site [https://www.nite.org.il/index.php/en/tests/psychometric/psychometric-structure.html]

The test's reliability, validity and fairness are routinely examined by NITE, to ensure the quality of the test (e.g., Kennet-Cohen, Bronner & Oren, 1999; Turvall, Bronner, Kennet-Cohen & Oren, 2008).

The universities use a simple average of the PET and the Bagrut to form an admissions score. Each department sets a different cutoff score to determine who would be admitted and the cutoff scores may change from year to year. Some departments may waive the PET requirement in order to attract more students. Academic colleges also require PET scores for admissions, while some colleges and the Open University do not.

In contrast to the matriculation tests, PET is viewed as a more standardized, more reliable and less sustainable to cheating. Yet, PET also suffers from negative publicity among the general population. Prospective students often think of the test as an unnecessary hurdle stopping them from getting their education and progressing in life. Part of PET's problematic reputation originates from an unexpected consequence– the thriving of preparatory institutes for PET. About 80% of PET examinees participate in a preparatory course, and some repeat the course (and the test) until they get a sufficient score. Although studies have shown the effect of coaching on test scores to be minimal and similar to the effect of studying alone (roughly 1/5 of the test's standard deviation, see Allalouf & Ben-Shakhar, 1998), most examinees feel that not taking a course would put them in a disadvantage. The courses are quite expensive, between $1,500 and $3,000, about 10 to 20 times higher than the cost of the test and almost equal to the university's tuition. The high course fees and the aggressive marketing strategies of these preparatory institutions contribute to the illustration of PET as an expensive exam and therefore, one that is biased in favor of richer populations.

The test is often portrayed by the media as flawed, biased, ineffective or otherwise redundant. Reporters often misinterpret the differential performance of various groups (gender, ethnicity, age) as indicating problems with the test, rather than portraying existing differences in Israeli sub-populations. Over the years, there have been multiple attempts by politicians to abolish or change the test. These attempts are often based on inaccurate, wrong or incomplete information. For example, in one bill to discontinue the test the following was written: "It is known that the tests scores are relative to the group of examinees tested in the same date. The scores are not absolute and therefore they hurt

the principle of equality." The test developers are well aware of the public dissatisfaction with the test, but for the most part, NITE's attempts to mend the test's public relations did not produce positive results. The research described below was motivated by the desire to learn more about the public opinion of the test, and use this information to improve the test and its image.

Method

Materials and Procedures

Data for this study was collected through online surveys, interviews and questions embedded in the feedback form filled out by examinees after completing the PET. This report focuses on the results from the online surveys and the feedback form. The interview data has not yet been analyzed.

A link to the online survey was sent by e-mail to past and future PET examinees, who provided their email address when they registered for the test. A link to the survey was also placed on NITE's website. In addition, the survey was sent to staff in IHE and educational organizations, members of the Israeli Psychometric Association (ISPA) and the employees of the National Institute of Testing and Evaluation (NITE), politicians, public figures and media representatives. Unfortunately, there were very few participants from the last three groups. Participation was voluntary and a prize of 200 shekels (approximately $50) was randomly awarded to ten survey participants from the student population.

Survey items were developed based on the FV mapping sentence (described above), previous items from PET's feedback forms, and a collection of arguments for and against PET (derived from news reports, journal articles, and bills for canceling the PET). The items were revised and refined through several cycles of pilot-testing and reviewing. The survey items focused on various issues related to the assumptions explicated in the interpretive argument for FV, such as: the relevance of PET's various domains and item types for academic studies, the extent to which PET fulfills its purpose, the test's reliability and fairness, the appropriateness of using PET scores and other instruments to

make admission decisions, the factors that can influence one's test score and the existence of alternative interpretations or uses of PET scores. Responses to most items involved choosing a single option from a set of increasing categories (e.g., responses to statements about PET were *don't know*, *strongly disagree*, *disagree*, *agree*, and *strongly agree*).

There were four open ended questions in the survey. One question asked "Have you encountered other uses of PET scores that are different than admissions to IHE?". Another question was directed only to IHE staff and asked "How does the PET help you in your work?". Another question provided an excerpt from NITE's website describing the purpose of the test: "The Psychometric Entrance Test (PET) is a tool for predicting academic performance, and is used by institutions of higher education to screen applicants for the various departments. The test ranks all applicants on a uniform scale and, compared to other admissions tools, is less affected by differences in applicants' backgrounds or other subjective factors." Respondents were then asked- "Do you believe that PET achieves its purpose?". Another question asked for any feedback, good or bad, about the PET, its purpose, its quality, or anything else related to the test.

Due to its length, the survey was divided into two parts. Each part contained questions about the characteristics of the respondent and a subset of the FV-related survey items. The open-ended questions were also divided between the two parts. Past and future examinees received a link to one of the survey parts and. About half the participants were given the link to the first part of the survey, and the rest received a link only to the other part. When participants completed the survey, they had the option of responding to the other part. Participants in other groups (NITE, IHE, etc.) received a shorter version of the survey that included the main questions from both surveys. Data collection took about four months. We will not report all the findings from the survey, only those that relate directly to the interpretative argument.

Responses to the feedback form were collected at the end of two PET administrations (October 2011 and December 2011). The PET's feedback form is not mandatory, but has a high response rate (around 80% for Hebrew-speaking examinees and 60% for Arabic-speaking). Due to space limitations we included only one item about FV and one item about selection instruments. A slightly longer version of the feedback

form was pilot-tested with prospective PET examinees who participated in a study on the possible inclusion of a writing task in PET (they completed a full practice exam before answering the feedback form).

Participants and Response Rate

Overall there were 11,665 respondents to the online survey. Of these, 10,630 originated from about 65,000 e-mails sent to people who registered to take the PET, which gives a response rate of about 16%. In addition, 766 respondents were directed to the survey from NITE's website. About 42% of the student respondents completed both parts of the survey. Other groups included 220 professionals (IHE staff, ISPA members and public figures) and 49 respondents were test developers and other employees from NITE. In addition, there were roughly 21,600 examinees who responded to the questions on the PET feedback form at the end of their exam (including 1,197 from the writing experiment).

Some of the results refer to the open ended questions in the surveys. We mainly focus on the analysis of the question concerning whether PET achieves its goals (*N*=6,938). These results are based on the content analysis of a randomly selected sample from the student group (past, present and future) and all of the responses from the IHE and NITE groups. In total, about 31% of the responses to this question were coded based on the topics that emerged in the response. The remaining responses have not yet been coded.

Background characteristics of the survey respondents

1. **Gender**: Roughly 57% of the survey respondents were female, which is about the same percentage of female PET examinees.
2. **Date of birth (year)**: The average respondent age was 23 (*SD*=6).
3. **Primary language**: 85% of respondents listed Hebrew as their primary language, 8% listed Russian, 6% listed Arabic, 4% listed English, and roughly 2% listed other languages. Respondents could list more than one language. The proportion of Arab students in Israeli IHE is about 11%, but almost a third of the PET examinees take the test in Arabic (the test is translated from Hebrew into 5

languages). Due to budget and time limitations, the online survey was only administered in Hebrew. This might have had a big impact on the ability of members of the Arabic-speaking population to participate in the study. Because the feedback form is given at the end of an actual exam, and it is translated into Arabic, that sample had a larger proportion of Arabic speaking examinees- about 22%.

4. **Occupation**: Respondents could select more than one option to describe what they do. 20% of the sample classified themselves as students at IHE and 18% classified themselves as high-school students or soldiers. More than 65% indicated that they currently work and 15% indicated they are unemployed. Of those who work, about 42% indicated work in temporary jobs, 33% indicated work in non-temporary jobs, 8% indicated work in public institutions (including education), 4% were self-employed and 1% indicated they were in senior positions (another 10% were classified as "other").
5. **Student Status:** Respondents could choose between (a) was a student in the past, (b) currently a student, (c) planning to be a student, or (d) never was and not planning to be a student. We compared this variable to the previous variable (Occupation) to classify respondents. Overall, 67% of the sample indicated that they were planning to become students, 26% were currently students. 7% were students in the past, including members from the professional group (roughly 2%). The 6% difference between the percentage of current students in the question about occupation and the 26% presented here could be attributed to uncertainties related to the wording of the questions and the timing of the survey. For example, most responses were collected during the summer so those who were about to graduate or were about to matriculate might have given contradicting responses about academic status. The remaining respondents included 213 from IHE, and measurement experts (from ISPA), 7 from the media and politics and 49 employees from NITE.
6. **Fields of study**: Because most respondents were future students, their responses indicated what they wish to study, rather than what they actually study. About 43% of respondents selected more than one field of study, so the median number

of fields was one. About 36% of the respondents were classified as studying (or having studied or planning to study) in Humanities-oriented fields, 47% in science-oriented fields, and 17% listed both. Table 1 lists the full breakdown of fields of study for all survey respondents. Note that respondents could select more than one field and the percentages are from the total sample so they do not add up to 100%. Table 1 cannot be directly compared to the distribution of fields chosen by examinees on the PET registration form (this information is available on the NITE website) because of the ability to choose multiple fields and because 32% of the survey sample included current or past students (as opposed to future students).

Table 1. Distribution of fields of study among the online survey respondents

Field	Percentage (N= 11,665)*
Engineering	23%
Economics, Business, Accounting	18%
Medicine, Pharmacy and Dental medicine	16%
Exact Sciences	15%
Social Sciences	13%
Psychology	12%
Computer Science	11%
Architecture, Design & Arts	11%
Education	10%
Para-medical professions & nursing	10%
Law	8%
Other fields	8%
Other Humanities-related fields	7%
Social work	4%
Agriculture	1%

*Respondents could select more than 1 option

7. **Latest PET score**: respondents could choose not to answer this question and we have no way of distinguishing between such potential individuals and those who did not take the test. Roughly 64% answered this question. We were able to identify the real PET score for 33% of those who answered the question. Overall, responses were quite accurate. The majority of people (63%) gave an accurate score, 31% gave higher scores than they actually got, and 6% gave lower scores. On average, the difference between the real and the reported scores was about 10 points (SD=47). Those who reported a higher score than they actually got

increased their score by 50 points on average, and those who reported a lower score than what they actually got decreased their score by about 50 points. The average real PET score for those who reported a lower score was 78 points higher than the average real score for those who reported a higher score. There could be several reasons for these findings, including self-enhancing biases, misremembering, typing mistakes, providing the highest rather than latest score, etc. Overall, the average survey respondent had a somewhat higher score than the average PET examinee. One could suspect that people with higher scores would tend to favor the test. In that sense, the results below may be considered a positively over-estimate of the true public opinion.

8. **Influence on opinion**: A subset of the respondents were asked to indicate which factors influenced their opinion about the test (*N*=6,375). They could choose more than one factor as their answer, and on average, each selected about 2 factors. About 96% of the sample chose either "my own experience" or "the experiences of those related to me", as the factors influencing their opinions. Other influences were the opinions of parents, teachers or professors (about 18%), and politicians or media figures (about 7%).

Results

In this section, we describe evidence concerning each of the assumptions regarding the public's perceptions that were laid out in the interpretive argument. The results we present here originate mostly from the various questions on the online surveys, but also from the feedback forms, and other sources of information. Next, we discuss each part of the interpretive argument and conclude with an overall evaluation of the test's FV.

Scoring 1: people think that the test scores are calculated accurately and consistently and appropriately represent performance on the test.

The scoring inference is based on beliefs that the score is an accurate representation of performance. A threat to this assumption can stem from beliefs that the way scores are calculated is unreliable or inaccurate. Note that if a person felt sick during

the exam, he or she may think their score does not represent their true ability, although it does represent their actual performance in these special circumstances. In other words, the scoring assumption does not relate to factors that influence performance, but rather to factors that influence how scores are calculated.

Respondents to the survey indicated they believe that PET is carefully and professionally constructed. About 72% of the respondents agreed with this statement (*N*=8,313). About 75% of the respondents agreed that it is harder to cheat on PET than on the Bagrut (*N*=8,354). These results suggest that most people think positively of the quality of the test and its administration. Such beliefs support the notion that it is unlikely to receive an incorrect test score. The customer relations division at NITE receives over 6,000 phone calls and emails each year. Very few of those refer to the way scores are calculated. On average, less than 1% of the examinees in any given year appeal their test scores. In most cases, examinees believe they should have received a higher score because they received higher scores on practice exams given by a preparatory institution. Since the practice exams are not necessarily real exams, are not administered in the same conditions as a real exam, and are not scored using the same calibration methods, it is perfectly reasonable that people would perform differently on them and on the real exam. In summary, only few people each year think their test scores are incorrect, and in most cases their reasoning has more to do with the practices of preparatory institutions than with the actual test. This suggests that for the most part people think the scores are an appropriate and accurate representation of their performance.

In the responses to the open ended question, the issue of how scores are calculated and used was frequently mentioned (in about 21% of the responses). Most criticism concerned the issue of the departmental cutoff scores or the weights given to the different PET domains. Although these issues are related to the test, they mostly reflect concerns with how the universities use the test scores, rather than how the test scores are calculated. The concern regarding the domain weights stems from the belief that the quantitative domain is less relevant to humanistic fields and the verbal domain is less relevant to scientific fields (we will elaborate on these perceptions later).

About 4% of the respondents argued that the scaling procedure is problematic. For the most part, these arguments stem from a misunderstanding of the purpose of

scaling and its implications. These respondents incorrectly thought that their scores are affected by the particular group that was tested with them on the same date. For example, one response claimed that the scores are not "personal" because they are relative to the rest of the group. Since this issue was infrequently mentioned, we can conclude that for the most part inferences related to scoring do not seem to have a face validity problem.

Generalization 1: people think that the test scores are based on a representative sample from the universe of observations.

The administration of PET is conducted in the most standardized settings possible within the practical constraints of the test. One issue that was raised in the past is whether there are fluctuations in the difficulty of the PET from one testing date to another. The difficulty of the test is kept at the same level in each administration by using extensive pilot testing and remedial statistical measures. Consequently, it should not matter on which date you take the test, your score should be very similar. However, due to differences in the population of test takers, the average PET scores differ from date to date. For example, February is usually when the brightest high school students take the test, with the goal of applying to IHE before their mandatory army service, through a selective military program. Consequently, grades on that date tend to be somewhat higher. Test preparatory institutions advise their clients not to take the test on that date. Of course, since grades are equated across all testing dates, it makes no difference when you take the test. Still, 40% of the respondents (N=8,156) agreed that "On certain administrations of the PET, it is harder to receive a high score", 27% did not know whether this statement is true or not and only 33% disagreed. This suggests that although NITE has consistently attempted to clarify this issue (on the PET's website, preparation materials and in the media), a large portion of the public still believes that the test's difficulty varies from date to date. Such a belief means they do not think that the test items are a representative sample from the universe of observations, at least not across different testing dates.

Generalization 2: people think that the number of items is large enough to provide a reliable estimate of the true score.

There were no specific questions about the length of the test and their relevance to its reliability. However, the length of the test is closely related to the time it takes to

complete it. Given that there are roughly 200 items on the PET and it takes about 3.5 hours to complete, the issue of time-related stress has long been a major concern of the examinees. For example, of the 8,337 respondents who answered the item, 78% agreed that "The PET's time restriction does not allow examinees to showcase their true abilities". Apparently, people perceive the length of the test not in terms of the reliable estimate one can obtain based on the items, but rather in terms of the effort required from the examinees. More items on the test means the test takes longer to complete and therefore examinees are expected to get tired and perform worse than they could. Alternatively, the stress involved in completing a long test also has an effect on the emotional and cognitive state of examinees and hence on their performance. In that sense, people believe that performance is differentially affected by the length of the test (time-wise and item-wise), and consequently, the test does not measure performance reliably. The issue of test length also surfaced in responses to the open-ended questions. Roughly 20% of the respondents considered the time/length issue as a problem that deters the PET from achieving its goal of providing a uniform, clear and relevant criterion for admissions. Interestingly, NITE has shown that additional time does increase test scores but does not change their relative order (Kennet-Cohen, Bronner, Oren & Eitan, 2008). This information has been made public and mentioned in the media whenever the issue was raised. Adding more time would essentially shift the score distribution to the right, which means that departmental cutoff scores would also increase, and the net effect on the individual decision outcomes would be zero.

Extrapolation 1: people think that the test items require abilities that are relevant for successful academic performance.

Extrapolation is the type of inference most related to the operational definition of FV. Consequently, there were multiple items concerning this issue. For example, respondents rated the relevance of each PET domain and the test as a whole to studies in Humanistic vs. Scientific-oriented fields. The response options were: *no relation*, *weak*, *mediocre*, *strong*, and *don't know*. Table 2 presents the distribution of responses to the question "in your opinion, what is the strength of relationship between the abilities tested in each of the PET's domains and the abilities needed to succeed in (humanities-oriented

or Science-oriented) academic studies?" A description of the typical fields in each orientation was listed below the question on the survey.

Table 2. Distribution of responses about the strength of relationship between abilities tested on PET and academic studies

Field	PET section	N	Don't know	None	Weak	Mediocre	Strong
Humanities oriented	Verbal	8,358	2%	8%	16%	32%	42%
	Quantitative	8,315	2%	25%	40%	24%	8%
	English	8,315	2%	6%	13%	36%	43%
	All	8,082	3%	11%	32%	45%	10%
Science oriented	Verbal	8,336	2%	19%	31%	34%	14%
	Quantitative	8,315	2%	6%	10%	24%	58%
	English	8,301	2%	7%	14%	31%	46%
	All	7,867	3%	9%	21%	42%	25%

There are several interesting points to infer from Table 2. First, most people have a firm opinion on the matter (the proportion of "don't know" responses was small and constant). Second, as expected, most respondents believed that the verbal domain is more relevant to humanities-oriented fields (74%) and that the quantitative domain is more relevant to science-oriented fields (82%). Third, the reverse pattern is stronger for science-oriented fields, that is, more people think the verbal domain is relevant to those fields (44%) than people who think the quantitative domain is relevant to humanities-oriented fields (32%). Fourth, English is perceived as important to both orientations (about 78%). Finally, the test as a whole is seen as slightly more relevant to science-oriented fields (67%) than humanities (55%). A closer inspection of the cross-tabulation of these questions (N=7,807) shows that roughly 25% of the respondents thought that PET is either irrelevant or has a weak relation to both orientations, and 49% thought PET is moderately or very relevant to both orientations. When disagreements occurred they were usually in favor of science-oriented fields. About 16% of the respondents thought PET is not relevant for humanities but it is relevant for science, whereas only 4% thought the opposite.

Participants were also asked about each type of item on the PET (including the proposed writing task). The survey included a link that participants could use to remind themselves of the various item types. Participants indicated whether the skills measured by each item type are important for succeeding in academic studies (regardless of the

chosen field of study). In addition, participants were asked to respond with respect to each domain in general. Responses were given on the following scale: *not important*, *somewhat important*, *very important* and *don't know*. The results are given in Table 3.

Table 3. Distribution of responses about whether the skills measured by different item types are important for academic studies

	Item type	N	Don't know	Not important	Somewhat important	Very important
Verbal	Vocabulary	8,134	1%	19%	45%	36%
	Analogies	8,111	3%	30%	45%	22%
	Sentence completion	8,065	2%	24%	45%	29%
	Root replacement	8,117	3%	56%	32%	9%
	Reasoning	8,095	2%	15%	35%	48%
	Reading comprehension	8,092	2%	3%	16%	79%
	Writing task (proposed)	8,090	10%	17%	33%	40%
	All Verbal	*8,089*	*2%*	*6%*	*49%*	*43%*
Quant	Quantitative problems & questions	8,101	2%	13%	43%	41%
	Quantitative comparison	8,090	3%	24%	45%	28%
	Inference from charts	8,104	1%	9%	30%	59%
	All Quantitative	*8,075*	*2%*	*8%*	*43%*	*47%*
English	Restatements	8,109	1%	19%	40%	39%
	Reading comprehension	8,109	1%	3%	16%	80%
	Sentence completion	8,119	1%	13%	42%	44%
	All English	*8,045*	*1%*	*4%*	*29%*	*66%*

Table 3 also shows some interesting results. First, most people have a firm opinion about the importance of the various skills measured by the items. The proportion of "don't know" responses was around 2% for all item types (with the exception of the writing task, which was largely unfamiliar to these respondents). Second, there are vast differences in the perceived importance of the skills measured by various item types, within and across domains. For example, while reading comprehension items are mostly seen as very important for academia (79%), root replacement items seem to be quite unimportant (only 9% rated them as very important). In fact, root replacement items and analogies are the two item types that respondents believed were of the least importance for academic studies. Reading comprehension (both in native language and in English) were seen as the most important, followed by inferences from charts, quantitative problems and reasoning. Finally, the respondents thought that each domain, as a whole, measures skills that are at least somewhat important for academic success. The

proportion of respondents who rated the Verbal, Quantitative and English domains as somewhat or very important were 92%, 90% and 95%, respectively.

The last set of evidence related to this inference comes from the PET's feedback form administered during two testing dates (Oct and Dec, 2011). Because this sample had a better representation of the Arabic-speaking examinees, we focus now on the differences between them and the Hebrew-speaking examinees, as shown in Figure 2. The figure shows the distribution of ratings in response to the same question as in Table 2. The proportion of missing responses was much greater for the Arabic-speaking examinees (about 30%) compared to the Hebrew-speaking examinees (about 9%). The proportion of missing responses varied across the items. The smallest sample (N_{Hebrew}=14,844, N_{Arabic}=3,986) was for the item referring to the whole test. All the missing responses were taken out of the analysis.

Figure 2. Distribution of responses about the relevance of PET domains to academic studies, and average PET scores by native language.

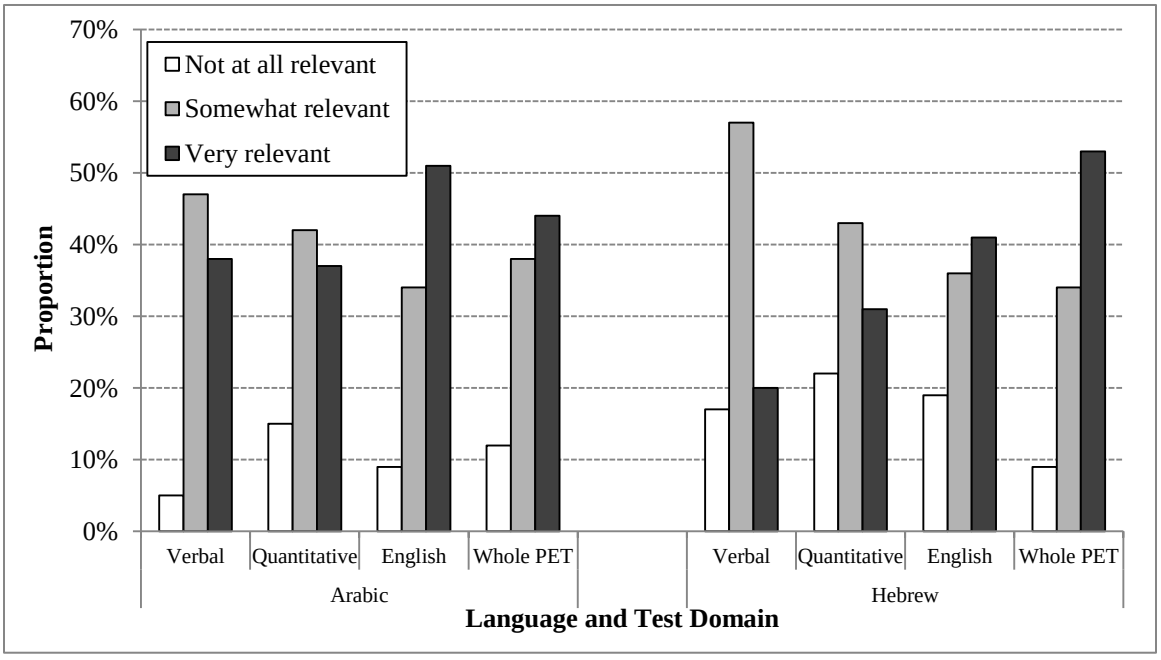

Figure 2 shows that for all domains, and PET as a whole, the majority of the examinees thought that relevant skills are measured. Overall, the verbal and quantitative domains seemed less relevant than the English domain. The majority of the examinees in both languages rated the skills measured in the whole test as very relevant to academic

studies. The proportion of examinees who thought the verbal domain measures relevant academic skills is much higher in the Arabic than Hebrew-speaking groups. This is surprising because most Arabic examinees take the test in their native language but their studies are conducted almost exclusively in Hebrew.

Extrapolation 2: people do not think there are other abilities that are relevant for successful academic performance but are not measured by the test.

It is important to clarify that the existence of important skills not measured by the test is not necessarily a problem. It would be naïve to assume that a single test can measure all the relevant skills needed to succeed as a student. The variability in fields and learning strategies is too wide to be captured by a single test. However, if some of the unmeasured skills seems crucial for admissions purposes, and can be measured reliably, then the test developers should consider adapting the test to include it. We measured perceptions regarding this extrapolation assumption, using direct and indirect items. The majority of respondents (90%) believed there are important skills that are not measured by PET. This seems to be true across all groups (students, IHE an NITE). It is obvious that there are other important skills, but it may be difficult to say what they are, whether they could be measured well, or whether they should be measured. Another item shows that most respondents (67%) thought that preparing for the test does not constitute a good preparation for academic studies. Although this item is reversely related to the extrapolation assumption, it does shed light on what people think about the test. If PET had measured all the relevant skills for academic studies, it is unlikely that people would have agreed with this statement.

Even if we know what should be measured, it is not always clear if indeed it is not measured by PET. Respondents were given a list of skills, abilities and qualities and were asked whether or not they consider them important for academic studies. If they said yes, they were asked if these skills were measured by PET. Table 4 shows the distribution of their responses. Respondents indicated several important abilities that are measured by PET: reading comprehension, time management, vocabulary, reasoning, speed thinking and speed reading.

Table 4. Distribution of responses to whether academically important abilities are measured by PET

	N	Don't know	Not important	Important but not measured	Important and measured
Reading Comprehension	8,322	1%	2%	9%	88%
Time Management	8,315	2%	6%	19%	74%
Vocabulary	8,304	2%	16%	9%	73%
Speed Thinking	8,328	3%	23%	6%	68%
Reasoning	8,316	2%	3%	26%	68%
Memory	8,330	2%	8%	26%	64%
Speed Reading	8,307	3%	37%	7%	53%
Persistence	8,316	3%	2%	52%	43%
Calculating without a calculator	8,318	3%	47%	9%	40%
Spatial Perception	8,305	13%	21%	32%	34%
Motivation	8,324	3%	2%	62%	33%
General Knowledge	8,316	3%	21%	57%	20%
Maturity	8,335	4%	9%	69%	17%
Creativity	8,304	3%	11%	71%	15%
Foreign Languages (not English)	8,332	5%	62%	23%	10%
Written expression	8,318	3%	7%	81%	9%
Curiosity	8,333	4%	14%	76%	6%

For some abilities, such as persistence, there was no consensus on whether it is measured by the PET or not, and on other abilities, such as manual calculation, there was no consensus on whether this ability is important for academic studies or not. Other abilities, such as motivation, general knowledge, maturity, creativity, curiosity and written expression, were identified as important but not measured by PET. The choices were constrained by the list in the item. Unfortunately, we did not ask participants to provide other abilities. When the issue of important abilities not measured by the PET was mentioned in responses to the open ended question about PET's goals (28% of the cases), the abilities indicated were mostly the ones that appear on the list.

Extrapolation 3: people think that there are no ability-irrelevant sources of variability that would bias the interpretation of scores as measures of ability level.

There could be many reasons why people think the test results are problematic. There are many factors that operate before and during the exam that can hinder the performance of some, and benefit the performance of others. Those who do not perform well on the test may be even more motivated to find faults in the test. Even those who do well on the test may think that part of their success is due to luck, prior knowledge of a particular domain or quality coaching and practicing.

As we mentioned before, most respondents thought it is harder to cheat on PET, which suggests people are not concerned with this ability-irrelevant source of variability. We inspected some of the most popular arguments about other factors that affect performance. Table 5 shows the distribution of responses to arguments about irrelevant factors affecting the test. In the first item, of the 8,337 responses, 78% said they either agreed or strongly agreed that "The PET's time restriction does not allow examinees to showcase their true abilities". This gives a strong indication that people generally do not think their performance showcases their abilities. The next two items show that over 70% of the sample thought the test discriminates against, or in favor of, particular sub-groups. Analysis of the open ended questions revealed that people thought richer examinees have better means for preparing for the test (e.g., they can pay for more expensive preparatory courses, take multiple courses, afford not to work while they study to the test), and therefore the test discriminates against the poorer examinees. In addition, people indicated that those who can think and act fast are more likely to succeed because they can better operate within the test's time constraints. Therefore, the test discriminates against those who work slower (e.g., immigrants, ethnic minorities, people with certain disabilities). The last item shows that a large portion of the examinees are not aware of the types of accommodations that are available to them or to other examinees. The remaining respondents did not have a consensus on whether sources of irrelevant variance are controlled for examinees with disabilities.

Table 5. Distribution of responses to items about ability-irrelevant sources of variation

	Statement	N	Don't know	Strongly disagree	disagree	agree	Strongly agree
1.	The PET's time restriction does not allow examinees to showcase their true abilities	8,337	1%	7%	14%	25%	53%
2.	PET is unfair in that it makes it difficult for applicants from weaker populations to be admitted to higher education	8,348	6%	7%	16%	30%	41%
3.	PET gives an advantage to people from certain socio-economic backgrounds	8,326	6%	12%	15%	30%	37%
4.	Examinees with disabilities receive accommodations that aid them in showcasing their true abilities	8,144	34%	16%	20%	22%	8%

In addition, about 50% of the responses to the open ended question about whether PET achieves its goals, indicated that performance on PET is influenced by irrelevant factors. Among the most frequently mentioned factors were: socio-economic background (30%), time pressure (18%), usage of test-wiseness or guessing (10%) and the quality of the preparatory course (8%).

Decision 1: people think that the conclusions about ability levels, as measured by the test, are meaningful for admission decisions.

This assumption essentially says that people believe there is a sufficient reason to use the test for admissions purposes. If the test provides meaningful information, and the conclusions about the examinees are warranted, the test should be used for admission purposes. Currently, the composite admission score is a simple average of rescaled PET and Bagrut scores. Respondents were also asked to indicate how they believed scores on the PET and the Bagrut should be weighted on the admissions score (*N*=7,558). About 37% thought the weights should stay equal. About 46% thought that the weight for PET should be reduced to .25, and 14% thought it should increase to .75. About 11% wanted admissions to be based only on Bagrut and 1% preferred it be based only on PET.

Another question asked respondents whether they agree that "PET provides a clear and uniform criterion for IHE in Israel". Only 32% agreed with this statement. This indicates that the majority of our sample thought PET was not providing meaningful information for IHE. Another result should be considered with respect to this finding. Only 48% of the respondents agreed that "the purpose of the PET score is to predict success in academic studies". This shows that about half the sample was either unaware or misinformed about the purpose of the score. Alternatively, disagreement with this statement may mean that people know this is what the score is used for, but they object to it being used that way. Another item can shed light on the issue. Respondents were asked to indicate what they believed the main goals of NITE were. Almost 60% thought that NITE's goal is to provide useful selection instruments for IHE. Taken together, these findings suggest that while people generally know what NITE aims to provide to the IHE, fewer people think that PET can be used in that way, and consequently, even fewer people think PET accomplishes its goals.

A related issue is the existence of other selection methods that may be used for admissions. We surveyed respondents about those and found that, on average, respondents chose three additional methods they thought should be used. The two most frequently selected methods were: admissions interviews (72%), and selection tests in the target field of study (70%). Another 47% suggested that final admission decisions should be based on first year's GPA, 32% advocated for personality tests, 31% selected using grades on relevant courses from pre-academic preparatory institutions, and 29% selected recommendation letters.

A similar question was included in the feedback form completed by PET examinees at the end of their exam (N=20,038). A list of selection instruments were presented and examinees were asked to indicate their appropriateness for admission purposes on a 4-point Likert scale, ranging from "not at all" to "very much". Again, the proportion of missing responses was much larger for the Arabic-speaking examinees (missing responses were taken out of the analysis). PET was rated among the least appropriate instruments, only 54% of the sample gave it high ratings (i.e., 3 or 4). The only methods that were rated lower were a writing essay and graphology. Again, the instruments that seemed most appropriate to the respondents were GPA on the first year of academic studies (89%), Bagrut (84%), GPA based on a year of studies in an academic preparatory institution (84%), interviews (82%) and recommendation letters (64%).

Decision 2: people think that students with low abilities, as measured by the test, are likely to perform worse academically than students with high abilities.

The purpose of the PET is to predict academic success. That is, for the most part, those who perform well on the test are expected to perform well academically. Multiple studies have shown the PET to have satisfactory predictive validity (e.g., Kennet-Cohen, Bronner, & Oren, 1995). Although this information is publicly available, people tend to argue against the test's ability to predict academic success. For example, over two thirds of the 8,345 respondents who answered the item, thought that "the ability of PET to predict academic success is negligible". This shows that years or research on the psychometric properties of the test have not helped in improving its standing in the public's eye. In responses to open ended questions, 35% of the answers indicated that PET does not achieve its purpose because its ability to predict academic success is

lacking. Only 6% thought that the score reflects chances for success. In most cases, respondents referred to themselves or to people they know, who succeeded in the academia even though their PET score was not very high, or vice versa. Obviously, focusing on single cases is not a valid way to evaluate the ability of PET to predict academic success. In conclusion, it seems that one of the strongest reasons to support the PET is misunderstood by or unknown to the public.

Decision 3: people think that using the conclusions about ability levels for admissions decisions does not have a negative social impact.

People may think that using the test has negative consequences for individuals or the society as a whole. For example, they may think that the time spent preparing for the test is a waste. Alternatively, people may think that using the test has positive consequences, for example, that the time used for studying to the test actually helps applicants get into the "student mindset". Thinking that the test has a positive impact does not imply that the test has no negative impact. Both perceptions can occur simultaneously. Yet, the existence of positive consequences may soften the existence of negative consequences.

Earlier we reported that the majority of people thought the test is unfairly biased against applicants from weaker populations. This is clearly one kind of a negative consequence of the test. Table 6 provides additional items. The first item in Table 6 shows that most people (63%) believe PET has more disadvantages than advantages.

One example of a negative impact is provided in item 2. About 73% of the sample thought that PET deters people from applying to IHE. Of course, this is only people's perceptions and may be completely false, but the fact that it is shared by so many respondents suggests that this is a concern. A selection test should help the universities select the best candidates. If people refrain from applying to IHE or settle on less selective institutions simply because they fear the test, this is worrisome. There is some indication that this problem is rooted deeper than the test itself. When asked about the main goals of NITE, 18% of the respondents thought it was to make it harder for people to get accepted to desired fields or institutions. This implies that NITE itself is seen as a gatekeeper, blocking the path to the most desirable academic institutions.

The third item shows that more than half the sample (60%) believed the preparatory course is necessary for obtaining a good score on the test. This can be understood as a negative consequence of the test because such beliefs boost the allure of the preparatory institutions. The high course fees (more than 10 times the price of PET) make the process of taking the test a significant financial expenditure.

The last item in Table 6 exemplifies a possible positive consequence of the test. Those who, for whatever reason, did not succeed on their Bagrut, may consider the PET favorably, as it provides them with an opportunity to overcome past mistakes and increase their chances of being admitted to academic studies. About 53% of the sample believed this is true about PET.

In response to the open ended question about the goals of PET, some people voluntarily mentioned various negative consequences of the test. For example, people argued that PET hinders their advancement and growth (13%), that PET is a waste of time and money (10%), that it leads to erroneous admission decisions (8%), and that it makes examinees feel frustrated, anxious, and disappointed (5%). When people mentioned the negative impact of PET on particular groups they often referred to poor and weak populations (19%), to people with disabilities (5%) or simply those who work slowly (5%). According to their arguments, the test is particularly difficult for those groups and therefore negatively impacts their advancement.

Table 6. Distribution of responses regarding the positive and negative consequences of the test.

	Statement	N	Don't know	Strongly disagree	disagree	agree	Strongly agree
1.	PET has more disadvantages than advantages as a selection tool for higher education	8,346	7%	7%	23%	28%	35%
2.	PET deters people from trying to apply to higher education	8,348	3%	6%	18%	34%	39%
3.	It is possible to succeed on the PET even without participating in a preparatory course	8,339	5%	26%	34%	25%	10%
4.	PET provides an opportunity for those who got low scores on the Bagrut	8,360	2%	21%	24%	37%	16%

Decision 4: people do not think conclusions about ability levels are used inappropriately for other purposes.

The PET scores are meant to be used by universities' admissions staff for calculating the composite admission scores, which are then compared to the stated cutoff point for each department. There may be other legitimate uses for the test's scores, although the test has never been validated for those purposes. For example, universities often accept students for their subsidized housing or fellowship programs based on their PET scores. Although the test scores are not supposed to be used in this way, it could perhaps be justified. Likewise, applicants for various jobs are often asked to provide their PET scores. This may be sensible for those who apply for an instructor position in one of the many preparatory institutions for PET, but it is probably not the best indicator of potential success in a big law firm.

To learn about such perceptions, one of the open ended questions asked "Have you encountered any uses of PET scores other than for admission to IHE?". Of the 8,371 participants who were given this question, only 49% provided a response. Of these, 65% said they had never encountered another use. However, the majority of these responses came from future students, some of whom had never taken the test. The remaining 35% indicated one or more other uses, such as:

Work-related (18%): Respondents reported that during job interviews people are often asked about their PET score, although many indicated they do not know if this actually has an effect on getting the job. The most frequently mentioned jobs were in high-tech companies, law firms and companies hiring accountants. Respondents frequently said that they (or their friends) listed their PET score on their CV. This practice is questionable because the applicants have already completed their undergraduate degree in the particular field of practice, and therefore their PET score is likely to represent irrelevant and archaic information about them. Often, respondents objected to using PET scores in this way. In addition, many mentioned that high scores on the PET are necessary for becoming instructors in PET preparatory courses.

Social (10%): Respondents mentioned the impact of the score on the individual and the social stigma that is associated with a low vs. a high score. They mentioned that the score is used to evaluate people's intelligence, unrightfully so. Many talked about the

tendency of high scorers (or their parents) to boast about their score. Others objected to the tendency of people to inquire about one's score in social settings. This issue was frequently mentioned also in response to the question about whether PET achieves its goal. It seems that many are offended by the way society classifies them based on this single number.

Academic-related (8%): Respondents mentioned the use of PET scores for certain programs such as one of the admissions criteria for the military academic project and the pre-academic preparatory institutions, criteria for receiving scholarships and fellowships, being admitted to advanced courses and honors programs in academia, and being admitted to on-campus student housing. Our own inquires on the topic revealed other inappropriate uses of the test in academic settings. For example, one college allowed applicants who did not meet the cutoff score to be admitted on probation. These students had to redo the PET during their undergraduate degree and achieve the desired score. This could create a paradoxical situation in which a student may graduate from the program with good grades, but would not get the degree because they have not yet received a sufficient score on the admissions test.

Conclusions

Summary of Current Theoretical Approach

In this paper we proposed that test developers and validators should consider additional sources of information when validating test scores. The additional information comes from the public's perceptions of the test, its goals, its structure, and the way the test scores are interpreted and used for various purposes. We argue that this information can be important for validity analysis because it highlights the alternative inferential networks at play in the reality of the test. It is crucial to examine whether the test is psychometrically valid, i.e., whether there is a valid chain of assumptions and inferences that lead from the examinee's performance to the decision about that examinee. However, in order to seriously test this chain, one has to consider not only what the test developers had in mind regarding the kind of inferences and uses of the test scores, but also how test users actually use the test, and how examinees and the public at large view the test and its appropriateness for the purpose it aims to achieve. Researchers who collect evidence

about the public's opinion of the test can identify whether the test's face validity is congruent or incongruent with the test's psychometric validity. That is, evidence about the face validity of test scores can be used to evaluate the appropriateness, accurateness and clarity of the validity of the intended uses and interpretations of test scores.

Moreover, public opinion can have a great impact on the test. This is especially true today when both information and disinformation can quickly and easily spread through the World Wide Web, and social networks are used more effectively to organize and put pressure on companies, decision makers and the behavior of individuals. The collection of evidence about the test's face validity does not only inform the validity analysis as explained above, but also provides test developers a way to gauge the level of acceptance of the test. Using this information, test developers can prepare for possible attacks on the test, consider ways to revise the test in a way that would make it more acceptable without hurting its psychometric qualities, or consider ways to improve the test's public relations to increase the dissemination of accurate and convincing information about issues about the test that seem problematic to the public.

In this paper, we presented a view that integrates the concept of face validity with the current framework of validity analysis. We suggested that a separate interpretive argument can be made concerning the face validity of test scores. We then presented how evidence collected through online surveys and other sources could be used to validate this argument. Note that a separate interpretive argument is not required. One could use the "regular" interpretive argument for test scores and simply add the perspective of face validity into each of the pieces of evidence collected to validate the argument. We did not discuss the possible advantages or disadvantages of each approach.

Limitations of the Study

The study described here has several limitations and should be viewed as a preliminary study that can lead to additional studies and developments in the future. The first limitation relates to the sample. Although there were many respondents to the survey, they represent a small portion of the relevant population. In addition, because the sample is based on self-selection, this may very well be a biased sample. Because we have found similar results in the responses to the feedback form (where the sample is more representative and less biased), we believe the issue of sample representativeness

did not significantly affect the results. However, future studies should be designed to better sample the relevant populations.

Another limitation has to do with the choices of items on the survey. We had designed the survey to match particular topics that were relevant to the interpretive argument in Figure 1, but there could be other issues that we have not included and are nevertheless relevant. Moreover, the specific wording of each item might have had an impact on the findings (e.g., if the item is worded negatively or positively). The open ended questions provided rich information that was less restricted to the topics we raised in the survey items. For the most part, the open-ended responses raised the same issues that we included in the items. Future studies should attempt to collect evidence that is relevant to the interpretative argument and at the same time, not as restricted as the list of items we selected.

Finally, the study helps us describe the status quo regarding the public opinion about the test, but it does not tell us how things might change. Future studies can be based on these findings to test various ways to deal with misconceptions about the test. For example, one can develop different ways to present facts and findings about the test, and test their effectiveness in reducing misconceptions and negative attitudes.

Summary of main findings and future directions

Coming into this study we already had a pretty good understanding of what we might find. Typical PET examinees are not afraid to voice their dissatisfactions or objections. We had known about many of the conceptions and misconceptions people hold about the test, what it measures, how scores are calculated and how they should be interpreted and used. This knowledge has been tacitly accumulating in the test developers' minds for years through formal and informal communication with examinees, such as: official complaints sent to the customer relations division, issues mentioned on the PET feedback forms, claims written in newspaper articles about PET, arguments raised by politicians in bills calling to change or cancel PET, and various discussions with examinees or organizations representing certain sub-groups of examinees. Most of these issues were studied by NITE in the various papers and technical reports discussed in the Introduction (see pages 12-13). There have been some indications of public opinion about PET (Nevo & Sfez, 1985), however, this is the first a large-scale study on the matter.

Scoring inferences

We have learned a great deal about the public's perceptions of the test. In many cases, the evidence we collected confirmed our informal expectations. We have found little evidence against the scoring assumptions. For the most part, people view test scores as an accurate, appropriate and consistent representation of their performance on the test. Exceptions are largely due to misunderstandings about the meaning of scaled scores and the deviation between scores on practice exams and scores on the real exam. Both of these issues could be alleviated by focusing public relations efforts to clarify these misconceptions.

Generalization inferences

The results show that there is less agreement in the public regarding the generalization assumptions. Many are unaware that the difficulty of the test is kept constant across administration dates, and therefore they are more prone to question the utility of the test scores. In addition, the psychometric advantages of a long test are overshadowed by the associated practical implications. Most people believe that the time stress involved in completing the test prevents examinees from showcasing their full potential, and therefore the test does not appropriately measure their ability. In other words, the relation of test length and the accurate measurement of ability are mostly not apparent to the public. Again, both of these misunderstandings could be addressed by the test's public relations. However, the issues concerning the length and time stress of the test do indeed pose a threat to its face validity. It is possible that addressing this issue in the future may significantly help in reducing the negative perceptions of the test.

Extrapolation inferences

The extrapolation assumptions are directly related to the traditional definition of face validity. We have found strong support that the verbal domain is perceived to be relevant to humanity-oriented fields, that the quantitative domain is perceived to be relevant to science-related fields and that the English domain is perceived to be relevant to both types of fields. At the same time, the test as a whole is perceived as more relevant to scientific fields. We have found evidence that Hebrew-speaking examinees think differently than Arabic-speaking examinees about the various components of the test.

Hebrew-speaking examinees tend to rate the relevance of PET domains lower, and the test as a whole higher, compared to Arabic-speaking examinees. We also found that people have different conceptions about the relevance of various item types for academic studies. Some items are seen as very relevant while others are seen as very irrelevant. Overall, the majority of people think that the PET as a whole, its domains and various item types, are at least somewhat relevant for academic studies. Our findings suggest that replacing certain item types with other items that seem more relevant may help improve the overall perceived appropriateness of the test for academia. Moreover, the findings suggest that using a different weighting scheme of the various domains for admissions to different departments is likely to improve the test's face validity.

On October 2012, a long-planned revision of the PET will become effective. Three major changes will take place: (a) traditional PET test scores will be reported to IHE along with two new scores, using a {.6, .2, .2} weighting scheme to emphasize verbal or quantitative abilities, (b) the time allocated for each section will be reduced from 25 to 20 minutes, and the number of items will be reduced accordingly in each section, (c) three item types will be discontinued (2 verbal- Vocabulary and Root replacement, and 1 quantitative- Comparisons), and (d) the first open-ended item in PET, a writing task, will be added to the test. These changes were planned long before this study began. It is reassuring to see that these decisions are supported by our findings.

The results show that the notion that other relevant abilities are not measured by PET is widely held by the public. The findings show several abilities that people think are relevant for academic studies, and there might be others that we have not asked about. Not all of these abilities can be easily or reliably measured, but some of them could probably be added to a future version of PET. Expending the scope of admissions criteria is probably a desirable approach although it involves significant efforts. Similar initiatives in other places (e.g., Sternberg et al., 2012) have produced promising results.

We have found strong evidence that people think there are ability-irrelevant factors that affect the performance on the test. Many believe the test is discriminating against weaker socio-economic groups such as immigrants, Arabic-speaking examinees, examinees with disabilities, and examinees coming from poor families. This issue has been raised multiple times in the past in various circumstances. NITE has shown that

much of the differences in performance can be attributed to actual differences in ability, which reflects variability in the quality of education. For example, the same differences in the performance of Arabic-speaking and Hebrew-speaking Israeli pupils can be found in international assessments such as PISA (Kennet-Cohen, Cohen & Oren, 2005). Studies on the fairness and differential predictive validity of PET have shown that the test does not discriminate against minorities, examinees with disabilities, or examinees from lower socio-economic status (e.g., Kennet-Cohen, 2001; Oren & Even, 2005; Turvall, Bronner, Kennet-Cohen & Oren, 2008). Apparently, the misunderstandings around this issue have not been resolved and are likely to surface again. Our findings suggest that these social and ethical questions cloud the public's discussion about the quality of PET as an instrument for educational measurement. We should think creatively about how to develop a constructive public discussion about these issues.

Decision inferences

The study results show that people often question the legitimacy of PET as a selection instrument for academic studies and think there are other instruments that could be used. Given that many advocate having admissions interviews, which are known to have very low predictive validity, it seems that the psychometric properties of the test do not impress the public. From their suggestions for additional selection criteria it seems that people wish they had an opportunity to express themselves, and that more field-specific abilities would be taken into consideration in their admissions decision. We believe that two of the forthcoming changes to PET are relevant steps in this direction (i.e., the addition of a writing task and the verbal vs. quantitative weighted test scores).

We have found strong evidence that most people are unaware or dismissive of the test's predictive validity. Most base their opinion on few personal examples and ignore the legitimacy of generalizing from these examples to the validity of the whole test. It would be difficult to address this issue systematically because this phenomenon is driven by internal psychological mechanisms (e.g., fundamental attribution error, availability heuristic) that are beyond the test developers' control. For example, one of the expected patterns, which was confirmed by preliminary analyses of these data (not reported here), show that the perceptions of the test are more positive for those who obtained higher PET scores.

The results also show that people think there are several negative consequences of the test. We had already discussed the issue of negative consequences concerning weaker populations. We believe these issues require a more comprehensive treatment, involving improvements to the resource allocation and quality of the education system in Israel. Overall, respondents tend to view the test as a hurdle preventing them from gaining their education. This may very well be an unavoidable attribute for a high-stakes selection test. Still, the negative implications of this view should be more actively addressed through the test's public relations. Our findings suggest that people are often unaware of the possible positive consequences of the test, such as giving a second chance for those who did not do well on the Bagrut or helping young adults get acclimated to academic settings (recall that most Israeli students begin their higher education about 3 years after they graduate from high-school).

Another negative consequence that was often mentioned involves the growth of expensive preparatory institutions. The common belief among the public is that a course is a necessary step for preparing to the test, and that more expensive courses would provide significantly better preparation. Although our studies have shown otherwise (Allalouf & Ben-Shakhar, 1998), these misconceptions persist. The unfortunate consequence is detrimental to the face validity of the test—if one can get a better grade simply by paying more for the course, then the test does not measure ability, but rather test-wiseness skills that can be bought, and thus are available mainly to the rich. These issues could be addressed in various ways- focusing public relations on this topic, adapting certain item types such that they would become less coachable, conducting new studies to gauge the true effect of preparatory institutions, and advocating legislation to regulate course fees or provide additional discounted courses without compromising their quality.

We have found some evidence that test scores are interpreted and used in various ways. Some of the interpretations and uses seem legitimate. For example, it may be reasonable to use the test scores as an additional source of information for determining who would get certain academic-related benefits such as scholarships, fellowships, subsidized housing, or admissions to special academic programs. It is probably not reasonable for people to choose a potential mate on a dating website based on some

cutoff for their PET scores. The most problematic aspects of these unintended interpretations are the ones related to the social and interpersonal arenas. In the responses to the open ended questions, people often argued against the test, claiming that "it does not measure intelligence". Indeed, PET is not an IQ test, so why do people think this is a problem?

The reason is that although they are not designed to measure intelligence, academic selection tests often correlate positively with IQ tests, and laypeople tend to conclude that high scores mean the people are smart and low scores mean they are not. This simplified view of the test has a detrimental effect on its face validity. Labeling people based on a single number is always unwarranted. People are afraid to get a low score, not only because they may not pass the cutoff score for their desired department, but also because they fear the social embarrassment involved. Consequently, people strive to pass a certain score (700 was often mentioned), even if it is much higher than what they need for admissions. This score inflation feeds into the perceived need to participate in expensive preparatory courses, and damages the true value of the test. Again, NITE has to think creatively about dealing with these misconceptions through research, public relations and outreach activities.

Final remarks

The findings show that the general validity framework is not necessarily hierarchical. For example, one might assume, based on Figure 1, that if X% of the people disagree with the hypotheses in one level of inference (e.g., generalization), then the proportion of people that disagree with the hypotheses in the subsequent levels of inference should be at least as high as X%. We have found that this is not the case. Moreover, within the same level of inference, people may tend to disagree with one hypothesis and agree with another. Even within the same hypothesis, people may have different perceptions about the various aspects of the test that relate to this hypothesis. Researchers conducting similar studies should design their methods and instruments such that they allow this variability to surface. In the analysis of such data, researchers should attempt to reconcile the competing evidence as much as possible.

One unexpected consequence of conducting this study was that many of the survey respondents had reacted positively to the survey. We had feared that because negative

views about PET are popular among the public, the survey would become an outlet for unrestrained attacks against the test. We had seen such responses, but they were not the modal type of response. Most respondents answered thoughtfully and respectfully even if they had very negative views of the test. Many respondents viewed NITE's decision to conduct such a study to be a sign of willingness to change or consider alternative points of view. They were happy to share their thoughts, and had done so in great detail in response to the open ended questions. Others suggested that we contact them for further communication, and had shown interest in learning about the results of the study. Overall, the experience of developing, conducting and analyzing this study was productive and constructive for NITE. The issue of public opinion is not only important for the face validity of the test, but also to the professional and personal lives of the test developers. We have used the setting of this study to foster discussions and initiatives internally within NITE. From these experiences we have learned that a more dynamic channel of communication between the test developers and the public may be informative, fruitful and generally beneficial for all.

References

Allalouf, A. & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests, *Journal of Educational Measurement, 35*, 31-47.

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

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: MacMillan.

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

Beller, M. (1994). Psychometric and Social Issues in Admissions to Israeli Universities. *Educational Measurement: Issues and Practice, 13,* 12-20.

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*, 315-337.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.). *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*, 327-333.

Downing, S. M. (2006). Face validity of assessments: faith-based interpretations or evidence-based science? *Medical Education*, *40*,7-8.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport. CT: American Council on Education/Praeger.

Kennet-Cohen, T., Bronner, S. & Oren, C. (1995). *A Meta-Analysis of the Predictive Validity of the Selection Process to Universities in Israel. Research Report No. 202.* Jerusalem: NITE.

Kennet-Cohen, T., Bronner, S. & Oren, C. (1999). *The Predictive Validity of the Components of the Process of Selection of Candidates for Higher Education in Israel. Research Report No. 264.* Jerusalem: NITE. [in Hebrew]

Kennet-Cohen, T. (2001). *Differential predication and differential validity of the higher education selection system based on the socio-economic status of the applicants. Research Report No. 285a.* Jerusalem: NITE. [in Hebrew]

Kennet-Cohen, T., Bronner, S., Oren, C. & Eitan, M. (2008). *The effect of extended time on scores and examinee rank order for three PET domains. Research Report No. 344.* Jerusalem: NITE. [in Hebrew]

Kennet-Cohen, T., Cohen, Y. & Oren, C. (2005). *Comparison of the performance of the Jewish and Arabic sectors on various levels of the educational system – a collection of findings. Research Report No. 327.* Jerusalem: NITE. [in Hebrew]

Lyrén, P.-E. (2009). *A perfect score. Validity Arguments for college admission tests*. (Unpublished doctoral dissertation). Umeå University, Umeå, Sweeden.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational & Psychological Measurement, 7*, 191–205.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22*, 287-293.

Nevo, B. & Sfez, J. (1985). Examinees' Feedback Questionnaires. *Assessment and Evaluation in Higher Education*, 10, 236-243.

Nunnally, J. (1967). *Psychometric Methods*, New York: McGraw Hill.

Oren, C. & Even, A. (2005). *The fairness and validity of the higher education selection system for students with disabilities*. *Research Report No. 325.* Jerusalem: NITE.

Oren, C., Kennet-Cohen, T. & Bronner, S. (2007). *Aggregated data about the validity of the higher education selection system for predicting academic success in the first year (the 2003-2005 cohorts). Research Report No. 342.* Jerusalem: NITE. [in Hebrew]

Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing.* New Brunswick, NJ, USA: Transaction Publishers.

Secolsky, C. (1987). On the Direct Measurement of Face Validity: A Comment on Nevo. *Journal of Educational Measurement*, *24*, 82-83.

Sternberg, R. J., Bonney, C. R., Gabora, L. & Merrifield, M. (2012). WICS: A Model for College and University Admissions. *Educational Psychologist, 47*, 30-41.

Stricker, L. J., Wilder, G. S., & Bridgeman, B. (2006). Test takers' attitudes and beliefs about the Graduate Management Admission Test. *International Journal of Testing, 5,* 255-268.

Turner, S. P. (1979). The concept of face validity. *Quality & Quantity, 13,* 85-90.

Turvall, E., Bronner, S., Kennet-Cohen, T. & Oren, C. (2008). *Fairness in the Higher Education Admissions Procedure: The Psychometric Entrance Test in Arabic. Research Report No. 349.* Jerusalem: NITE.