A Preoperative Index for Construct Validity

Ali Baykal

Bahcesehir University, Istanbul, Turkiye

ali.baykal@bahcesehir.edu.tr

Abstract

Judgmental procedures fall short of restoring the desirable attributes of the test after it has been administered. Preventive strategies must replace negating efforts. Perfect key reliability for instance can always be ensured during the construction of the test before using it. Apparently, attributes such as inter-subject reliability, predictive validity, concurrent-validity etc. cannot be predicted before obtaining the empirical data. So far as the construct validity is concerned there are some a priori aspects independent of responses given by the participants. The inclusion error (irrelevant impurities diffused into the items) can be detected and removed before the subjects are exposed to the test. Also exclusion error (misrepresented intent) can be identified and essential content can be supplied in advance. In order to describe the relevancy between the intent of the test maker and the effect as distinguished by the expert(s) a numerical index based on the Shannon's concept of entropy is introduced in this study. Items in the instrument are tallied into categories (taxonomical levels, sub-constructs etc.) as intended by the test-maker. Same set of items are checked in categories as distinguished by the expert(s). Observed frequencies are cross-tabulated on a contingency table to compute entropy values.

Keywords: Construct Validity, Content Validity, inter-subject reliability; information theory

Introduction

An assessment procedure departs from a reason whether practical or intellectual. There are so many compelling forces to make educational or psychological decisions. Selection and placement, guidance and counseling, formative evaluation, certification, program evaluation are the immediate examples of major purposes for testing. These aims are the determinants of testing practice, and they are also criteria for the accountability of the whole critical conduct (Hambletone &Zaal, 1989; Shepard, 1993).

Theoretically there are three major desirable attributes expected of instruments and methods involved in the measurement procedure: Validity, reliability and practicality. There are also different aspects of these triumvirate qualities. There are almost two dozen types of validity. There are at least five types of reliability: Key reliability, intra-scorer reliability, inter-scorer reliability, intra-subject reliability and inter-subject reliability. There are at least nine aspects of "practicality" when the three phases of testing (i.e. preparation, administration, reporting) were cross-tabulated with respect to criteria (i.e. cost, ease, time needed). These qualities are not independent accessories assembled arbitrarily; their merits and drawbacks are interdependent. They cannot all be maximized at the same time for all kinds of purposes. Even if their descriptive definitions were the same their prescriptive magnitudes and directions would be case-specific. They have to be optimized to fulfill the requirements of a particular precise purpose of evaluation. In a large-scale, high-stake testing objectivity (scorer reliability) could be the foremost

necessity. For example in a country where the corruption perception is high face validity of open ended exams will tend to be low. Machine scoring will definitely uplift the scorer reliability in measuring convergent abilities and mental achievement at all taxonomical levels but synthesis. If the test fails to cover creativity, synthesis, and other similar divergent skills then the content validity will be low. Chance success is an inevitable impurity in educational achievement as measured by choice type of exams. Inclusion of undesirable constructs means lower construct validity. In selection and placement exams predictive validity is more favorable than the content validity which cannot be sacrificed in curriculum evaluation (Cronbach, 1971).

The Need to Quantify Construct Validity

Measurement is a procedure which requires instruments and operations. In physical measurements ratio or interval level quantity qualifies the quality of the construct. Tests are the most common measuring instruments in behavioral sciences. In its general sense, validity is the relevancy, consistency, compliance, concordance, conformity between what is intended to be measured and what is really to be measured. Some hypothetical examples are as follows: One wants to measure intelligence but factual information is examined; everybody favors creativity to be tested but what is actually being measured can be cognitive achievement. Figure 1 illustrates the construct validity.

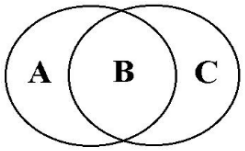

Figure 1: relevancy between what is intended to be measured and what is really measured

In Figure 1 area A+B is the set of elements that are purported to be measured. Area B+C is the set of elements that have been measured in reality. In other words area A involves elements which couldn't have been measured although they had been intended to. Zone C includes the impurities. Zone B is the extent to which test measures what it intends to measure (construct validity). Table 1 displays relevant and irrelevant examples in a selection and placement test. Cell D corresponds to the universal set which is implicit in Figure 1 (Ferrara, 2007).

Table 1: Summative Analysis of Construct Validity in a Selection Exam

		Outcome	
Intention		Not measured	Measured
	Desired	C: Analysis, Synthesis, Creativity	B: Knowledge, Comprehension
	Undesired	D: Weight, Height, Eye color	A: Chance success, Income, Anxiety

One ought to admit that there is no verbal definition of any quality better than an index which can express it in terms of a quantity. The conceptual definition of validity yields two implications. First, construct validity is a matter of degree therefore it is itself a construct to be measured. Second, validity is not independent of the purpose, isolated from the intent, sovereign to the expectations of the experimenter. Construct validity tells the experimenter the extent to which test behaviors correspond with the constructs defined by the theory. Messick asserts that "…**construct validity may not be the whole of validity, but it is surely the heart of it.**." (Messick, 1989: 2)

Paradoxes omitted for many years

In testing human traits the zero point is set differently for every single participant. Also total scores obtained from tests or questionnaires are not composed of uniform, equidistant units. By ignoring the error of isomorphism a myriad of advanced quantitative methods have been developed to extract information embedded in the response data. These methods yield some numerical indicators to delineate some important characteristics of items and of the test.

These empirical quantifiers may suggest removal of non-discriminatory or low-variance items from the test. Whereas every single item assembled into the test must have been manufactured by hard mental work (Osterlind, 1989). Such curative procedures fall short of restoring the desirable attributes of the test after having been administered. Shortly preventive strategies must replace corrective efforts. Perfect key reliability for instance can always be ensured during the construction of the test before using it. Some empirical attributes such as inter-subject reliability, predictive validity, concurrent-validity etc. cannot be predicted or estimated without having the real data from in vivo practice. There so many a priori aspects of construct validity and also content validity independent of responses given by the participants. Therefore there is a possibility to make it straight at the very beginning. This possibility entails the responsibility to do so.

The importance of ensuring test validity is obvious. In its general sense, validity is the relevancy, *consistency, compliance, concordance, conformity between what is intended to be measured and what is really to be measured* (Cronbach, 1971). One ought to admit that there is no verbal definition of any quality better than an index which can express it in terms of a quantity.

At present, construct validity of tests are reported based upon; 1. expert opinion; 2. content analysis; 3. factor analysis; 4. correlational analysis (concurrent validity).

First two of these methods have to be applied and ensured before the test is administered. They may not be objective, parsimonious and informative. For the time being they cannot be reduced into a single numerical index which enables the users to

compare rival instruments. Also, these are usually a priori judgements about the test. The empirical findings may not be in compliance with the expectations. Factor analysis is a very useful descriptive tool but it is free from the prescriptive paradigm. As long as, there is no reference to manifest intent one can neither accept nor challenge the extent the appropriateness of the factor structure latent in the test.

Criterion referenced validity cannot be taken as a particular type of validity but a method to verify a certain type of validity. The correlation between the present practice and the criterion;

i. in the past may (or may not) be an indicator of construct validity;
ii. at the present may (or may not) imply concurrent validity;
iii. in the future may (or may not) verify predictive/consequential validity.

Correlation is a very useful tool invented in statistics. It can be used to quantify a lot of qualities, but one has to be cautious about its side effects. Attenuation, spurious correlation, the effect of extreme scores and especially the effect of combined groups are some of the examples to mention (McCall, 1975; McMillan & Schumacher, 2010).

Correlational analysis of construct validity is subject to all of these symptoms in general. In particular correlational analysis contradicts the conceptual definition of construct validity. Its frame of reference is always the previous practice. To the extent that the new test is in absolute conformity with the previous one how one can conclude about its uniqueness. Therefore, reporting a perfect correlation between the new test and the old criterion is nothing but a declaration of the fact that the new one has no originality or novelty over the old one. On the other hand, how can the validity of the old one be assessed? This task will be accomplished by looking at the correlation with the one older than the old one. How about the validity of the oldest of all then?

Negative correlations would be another source of difficulty for understanding the relationships between constructs. What would it mean to have obtained a negative correlation between the construct being tested and the one taken as the criterion?

To sum up, a numerical index other than correlation to reduce validity data into an informative criterion seems to be necessary. The purpose of the present study is to introduce a new quantifier for construct validity i.e. a new index for the construct validity of a new test for which even if there is no equivalent external criterion.

Proposal

In information theory the tendency in a system to proceed towards a state of greater disorder is expressed by the concept of entropy. When the system becomes more and more disorganized, one is less informed than before. So far as construct validity is concerned two courses of action can be defined: The first one is the “intention” made by

the test constructor (message sent). The other could be the expert opinion (message received). In order to describe the relevancy between the intent of the test maker and the effect as distinguished by the expert(s) a numerical index based on the Shannon's concept of entropy is utilized in this study. Items in the instrument are tallied into categories (taxonomical levels, sub-constructs etc.) as intended by the test-maker. Same set of items are checked in categories as distinguished by the expert(s). Observed frequencies are cross-tabulated on a contingency table to compute entropy values. The pre-operative construct validity of the instrument is defined as the uncertainty removed by the observed distribution over the total uncertainty observed in the distribution.

Preoperative Construct Validity can be measured starting from the point where the choices of the experts are completely independent of the sub-dimensions pre-set by the test maker i.e. in terms of the decreasing uncertainty departing from the maximum possible depending upon the number of options defined for the referees (experts). "a" is the number of sub-dimensions (factors, subscales etc. foreseen by the test maker(s). In a free format questionnaire the number of choices by the experts can be more or less than "a". alternatives which can be chosen plus 1 refers to all the other possibilities put together e.g. omissions, double choices etc. Test-retest responses of every single individual can be plotted on a Contingency Table as shown in Table. 2 below. In a structured questionnaire it can be (a+1) alternative responses. In this table, $f(x_i,y_i)$ stands for the frequency of items observed for the i^{th} sub-dimension "intended" by the test constructor corresponding to the j^{th} sub-dimension "perceived" by the referee.(s). The marginal total $f(x)$ represents the frequency of sub-dimensions intended by the test maker. Similarly, $f(y)$ is the frequency of sub-dimensions as "perceived" by the expert(s).

Table 2. Contingency Table for intended vs. perceived sub-dimensions

		Sub-dimensions "perceived" by the referee(s)						
		A	B	C	D	E	F	Total $f(x)$
Sub-dimensions intended by the test constructor	A	$f(x_1,y_1)$						
	B							
	C			$f(x_3,y_3)$				
	D							$f(x_4)$
	E						$f(x_5,y_6)$	
	Total f(y)					$f(y_5)$		Grand Total

These are defined by the formulas (1) and (2) respectively. The number of options is "a" for the test-maker, but judges are free to make choices more or less than "a". Here it has been taken as (a+1). The last option is "Another".

$$\text{Total frequency of intended subdimensions of the "construct"} : f\left(x_i\right) = \sum_{j=1}^{a} f\left(x_i, y_j\right) \quad (i = 1,2,\ldots,a) \qquad (1)$$

Total frequency of perceived sub-dimensions of the "construct": $f\left(y_j\right)=\sum_{i=1}^{a+1} f\left(x_i, y_j\right) \quad (j=1,2,\ldots,a+1)$ (2)

Shannon's entropy formulas to measure uncertainty in information exchange

Shannon & Weaver (1949) defined entropy as a quantitative measure of "noise" in a two way communication experiment. It has been applied in psychology (McGill, 1954; Attneave, 1959). Here, sub-dimensions of the test "intended" by the test constructor A, B, …, F choices correspond to signals sent. Respective choices of the referees correspond to the signals "perceived". The joint entropy measures how much uncertainty is enclosed within the cross-tabulated "intention vs. outcome matrix.

"K" number of items are cross-tabulated within a [a X (a+1)] contingency table. The agreement between intended & perceived responses corresponds to mutual information which quantifies the conformity between the intention of the test maker and their confirmation by a source of authority (e.g. experts). If the test and retest responses are completely independent the uncertainty will be maximum which denotes absolute absence of relevancy between the construct defined and the construct professed. When there is a perfect match between the construct structure declared by the test maker and the pattern observed by the expert(s) the mutual information will be maximum that implies perfect construct validity.

The uncertainty "designed" by the test constructor: $H\left(X\right)=\sum_{j=1}^{a} p\left(x_i\right) \quad (i=1,2,\ldots,a)$ (3)

The uncertaity "resolved" by the referent(s): $H\left(Y\right)=\sum_{j=1}^{a+1} p\left(y_j\right) \quad (j=1,2,\ldots,a+1)$ (4)

The joint uncertainty between "intended" and "achieved": $H\left(X,Y\right)=\sum_{i=1}^{a}\sum_{j=1}^{a+1} p\left(x_i, y_j\right)\ln p\left(x_i, y_j\right)$ (5)

The proportion observed/obtained in any cell of the matrix: $p\left(z\right)=f\left(z\right)/K$ (6)

Consistency between the "intended" and the "perceived" : $I\left(X,Y\right)=H\left(X\right)+H\left(Y\right)-H\left(X,Y\right)$ (7)

Where *K* is the total number of items presented by the test maker/judged by the experts.

By using these uncertainty (entropy) measures the preoperative index for construct validity is defined below:

Proposed preoperative index for construct validity: $g=I(X,Y)/H(X,Y)$ (8)

Practice

The pre-operative construct validity of the instrument is defined as the uncertainty removed by the observed distribution over the total uncertainty observed in the distribution. The implications of the index have been demonstrated on a questionnaire named "Personal Predispositions Perceived (PPP)". PPP uses 9 point Likert scale. There are 80 items representing all sub-constructs equally. Items are judged by 49 experts independently and completed by iterative use of the proposed procedure.

Procedure

Step 1. The questionnaire is based on the assumption that As a reflective practitioner the teacher ultimately is a decision maker. Regardless of their profession all decision makers may diverge in terms of their personality relevant tendencies while making their choices. In generating their ideas all people may assume different modes of thinking. Since 1997 colored hats metaphor has been being denoted in classifying ways of thinking. Edward de Bono (1999) identifies 6 hats in different colors: The White Hat calls for objective information, scientific reasoning. The Red Hat signifies intuitive thinking. The Yellow Hat stands for affective, moral approach. The Black Hat symbolizes critical, skeptical aspects of phenomena. The Green Hat implies opportunities, new ways and means, namely creativity. Finally The Blue Hat consolidates all the others by taking the best of each. As a matter of fact teacher educators concentrate nearly all of their efforts on competencies such as problem solving, critical thinking, and creativity. They also try to escalate the value of emotional intelligence, synergy at work, empathy, and optimism. Social partners and stake holders on the other hand expect of sociability and collaboration. Hard work, dynamism and action are the requirements of employers from teachers. Having taken for granted all these assertions made so far an 80 item Likert questionnaire has been prepared to assess the participants' mode of thinking as declared by them. "*Perceived Personal Predispositions*" is assigned as the name for the questionnaire. There are 8 sub-constructs involved: 6 of them are the ones signified by colored hats. The other two are the sociability and the dynamism. There are two reasons why these are included in the questionnaire. Firstly these two are also personality relevant attributes expected of people within any social context. Secondly these two are supposed to function as suppressors to mask the main six attributes aforementioned.

Step 2. The items were randomly sequenced and sent to 49 judges (experts). They were requested to identify the sub-dimension of each item from among the 8 options available. It could have been 9 to avoid rigid restriction. Instead the experts were encouraged to supply their free format comments, objections and suggestions. These open ended

remarks have been utilized in later stages of test improvement. Their choices are collected and counted. Each item is classified to the sub-dimension according to the most frequent vote taken from the experts. In other words "***Sub-dimensions "perceived" by the referee(s)***" in Table 2 have been identified according to the votes given by the experts. Table 3 has been prepared as a concrete example of Table 2 for the questionnaire mentioned above.

Table 3. Distribution of 80 items in the PPP into the sub-dimensions intended by the test-maker versus as perceived by the 49 experts

		49 EXPERTS' PERCEPTION								
		A	B	C	D	E	F	G	H	Total
TESTMAKER's INTENTION	A	9	0	0	0	0	0	0	1	10
	B	0	9	0	1	0	0	0	0	10
	C	0	0	4	1	4	1	0	0	10
	D	1	0	1	8	0	0	0	0	10
	E	0	0	2	0	8	0	0	0	10
	F	0	0	2	0	0	3	0	5	10
	G	0	3	1	1	0	0	5	0	10
	H	1	0	1	0	0	0	0	8	10
	Total	11	12	11	11	12	4	5	14	80

Step 3. Table 4 is prepared to compute the Shannon's "entropy" formulas.

Table 4. Entropy values of frequencies obtained for Table 3.

		49 EXPERTS' PERCEPTION								
TESTMAKER's INTENTION		A	B	C	D	E	F	G	H	T
	A	0.25							0.05	0.26
	B		0.25		0.05					0.26
	C			0.15	0.05	0.15	0.05			0.26
	D	0.05		0.05	0.23					0.26
	E			0.09		0.23				0.26
	F			0.09			0.12		0.17	0.26
	G		0.12	0.05	0.05			0.17		0.26
	H	0.05		0.05					0.23	0.26
	Total	0.27	0.28	0.27	0.27	0.28	0.15	0.17	0.31	

H(X) = 2.08 nit H(Y) = 2.02 nit H(X,Y) = 2.81 nit I(X,Y) = 1.29 nit

The units of uncertainties (entropies) are in "nit (Natural digIT) when computed in by natural logarithm. If logarithms were computed based on 10 the units would be in Hartley's.

Step 4. Finally the proposed preoperative index for construct validity has been calculated and found to be **g = 0.46.**

Its significance can be tested by benchmarking it with Pearson Chi Square, Lamda, Guttman and Kruskal tau. Computing entropy based Construct Validity index may not seem to be more practical in comparison with simple counts for Chi-Square computations. What remains is the intellectual value of theoretical coherence. There are some other entropy based indicators which can be used to quantify item and subject characteristics (Maccia, 1963; Hintikka & Suppes, 1970; Guiaşu, 1977).

Step 5. The main theme of this paper starts from here on. By using the frequency counts in Table 3 the items which did not appear along the diagonal were analyzed in terms of their cultural, verbal, psychological, educational connotations. A smaller team of colleagues and friends helped the researcher to improve the questionnaire.

Questionnaire has been administered on-line to 2149 teachers and education specialists working at İstanbul schools for the first time in 2010. Reliable, valid and interesting results have been obtained. Total size of sample reached to larger than 3000 participants from different type of institutions.

Acknowledgements

This paper is a by-product of the research project titled *Using Entropy Indices in the Analysis of Test Data* encoded as 09D6038. It has been supported by Bogazici University Scientific Research Projects Fund. The author wishes to express his gratitude to Administrative Coordination Office for Research Projects for their helping hand.

References

Attneave, F. (1959). *Applications of information theory to psychology*. New York: Henry Holt and Company.

Cronbach, L, J. (1971), Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 443-507). Washington, DC; American Council on Education.

De Bono, E. (1999). Alti Sapkali Dusunme Teknigi. 2. Basim. (Ceviren: E. Tuzcular). Istanbul: Remzi Kitabevi.

Ferrara, S. (2007). Educational Measurement: Issues and Practice Toward a Psychology of Large-Scale Educational Achievement Testing: Some Features and Capabilities, Springer London, 2007.

Guiaşu, S. (1977). *Information theory with applications*. New York: McGraw-Hill.

Hambleton, R. K., Zaal, J.N. (1989). Advances in Educational and Psychological Testing. Boston: Kluwer Academic Publishers.

Hintikka, J., P. Suppes, (Eds.). *Information and inference*. Dordrecht, Holland: D. Reidel Pub. Co., 1970.

Maccia, E.S. (1963). An Educational Theory Model: Information Theory. *Occasional Paper 63-141.* Ohio: Bureau of Educational Research and Service,

McCall, R.B. (1975). *Fundamental statistics for psychology. (Second Edition)*. New York: Harcourt Brace Jovanovich, Inc.

McGill, W.J. (June 1954). Multivariate information transmission. *Psychometrika.* 19:2, 97-116.

McMillan, J.H.; Schumacher, S. (2010). Research in Education (Seventh Edition). Boston: Pearson.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. Research Report. Educational Testing Service Princeton, New Jersey May 1981 RR-8i-9
Osterlind, S. J. (1989). Constructing Test Items. Boston: Kluwer Academic Publishers.
Shannon, C.E., Weaver, W. (1949). *The mathematical theory of communication*. Urbana: The University of Illinois Press.
Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education.* Vol. 19(1993); 405-450.