On testing plausible threats to construct validity
Schouwstra, S.J.

Citation for published version (APA):
CHAPTER 1.

THE CONCEPT OF TEST-VALIDITY:
A SHORT HISTORICAL OVERVIEW

Abstract. - Late in the 19th century quantitative measurement of mental phenomena started. Early in the 20th century the question of test-validity was raised. That first period (1925-1945) of the validity discussion was characterized by considerable consensus. Validity referred to the quality of a test, which could be established empirically and expressed in a single coefficient. After 1945 the consensus, regarding the procedure of establishing validity, started to disappear. Different kinds of validation procedures were distinguished depending on the particular aim of testing. In the early sixties the validity discussion died down, because once again consensus was achieved. Three different aims of testing and three kinds of validity were distinguished. In the next twenty years the pragmatic approach to validity was abandoned and a more philosophic approach emerged. The reasoning based upon test-scores became the subject of validity inquiry, rather than the quality of a test. As a consequence of the more philosophic approach, testing theory and testing practice became detached. After 1980 the validity discussion revived. The increased test use had resulted in a public debate concerning the fairness and consequences of test use. Furthermore, the relation between validation theory and validation practice was tense and discrepant. As a consequence the validity issue was expanded: test value, the consequences of test use, and test development were incorporated into the definition. Although in theory, validity now refers to everything that comes into play with testing, no significant cohesion between validation theory and validation practice has emerged.

INTRODUCTION

As far back as written history takes us, we can find examples of people being put through tests of mental capacity. The use of written tests takes us back to a very distant past in China, but Galton (1822–1911) is given credit for the parentage of quantitative measurement. In the 19th century the scientific development and use of so called mental tests started, but it was not until 1914 that, for the first time, the issue of the validity of these tests (Freeman, 1914) was a main topic in an article (PsychLit). Prior to 1914, validity was predominantly associated with psychophysiological laws and methods (such as reaction time) for measuring mental phenomena. In those early years of
psychological testing the concept of validity was assumed to be clear and seen as a test property that could be established. But quite soon problems were encountered, and the concept of validity and validation themselves became subject of scientific inquiry.

1925-1945: YEARS OF CONSENSUS

In 1927 Kelly concluded that validity is one of the most difficult problems facing a test deviser.

"The establishment of the fact that a given test is valid for a specifically named purpose is at present one of the most, if not in fact the most, difficult of the problems confronting the test deviser." (p. 30-31)

He defined the problem of validity as the problem whether a test really measures what it purports to measure (p.14), which remained the definition of validity until the early fifties (e.g. Adkins, 1947, p. 160; Garrett, 1937, p. 324; Hunt, 1937, p. 20; Turney, 1934). Until 1920 the establishment of validity rested predominantly upon expert opinion. But Kelley viewed expert opinion ("those competent to judge") only as an initial form of validity evidence. He stated (p. 30):

"The correlation between a test proposed as one having prognostic value (and what test is not so proposed?) and later demonstrated degrees of success or failure is the final measure of whether the test is actually valid for the purpose claimed."

Until mid forties the correlation of the test with what is later called a criterion\(^1\) remained the measure of validity or even becomes equated with validity itself. Bingham, for example, defined in 1937 validity in operational terms, in accordance with the behavioralistic view of science, current in that time:

"Validity is the closeness of agreement between the scores and some other objective measure of that which the test is used to measure."

But finding such an objective measure (criterion) proved to be difficult and other measures or procedures of validation were allowed. For example Garrett (1937) remarked that, if a reliable criterion was unavailable, the validity could be determined by indirect methods. He mentioned two indirect validation

---

\(^1\) A criterion is an objective measure in terms of which the value of the test is estimated or judged. " (Garrett, 1937)
methods: the correlation of the test with other tests, designed to measure the same function (what he calls the representativeness of the test), and, second, the judgement of experts as to the suitability of the material included.

1945-1960: VALIDITY AND TEST PURPOSE

The period of consensus with respect to the definition of validity and procedures to establish validity started to crumble in the second half of the forties. The concept of validity became the focus of many articles and more emphasis was placed upon the purpose of the test. For example, Adkins (1947, p. 160) stated:

"The test must be valid for the purpose for which it is to be used."

This emphasis was not very surprising given that the difficulties with finding a reliable criterion were specifically prominent with certain types of tests (e.g. personality tests). Subsequently, the so-called indirect methods of validation were the sole methods employed within certain domains of psychology.

In align with the different procedures employed for establishing validity, different kinds of validity were proposed. For example, face validity (Mosier, 1947), factorial and practical validity (Guilford, 1946), and logical and empirical validity (Cronbach, 1949). These distinctions explicated the idea that a test might be of worth (valid) even when the test proved to be a poor predictor, if that test was not intended as a predictor of performance.

Due to the many distinctions made, the American Psychological Association perceived it necessary to clarify the validity issue, and appointed a committee. In 1954 the work of this committee was published. In that publication, validity was defined as the degree to which the test was capable of achieving certain aims. Four different aims of testing were differentiated and four types of validity were distinguished: construct validity, content validity, predictive validity and concurrent validity (see Table 1.1).

New in this list is the concept of construct validity. Cronbach and Meehl (1955) stated that construct validity had to be investigated whenever no criterion or universe of content was accepted as entirely adequate to define the quality to be measured (p. 282). Construct validity was introduced specifically for tests designed to measure constructs\(^2\) that could not be operationally

\(^{2}\) A construct is some postulated attribute of people, assumed to be reflected in test performance (Cronbach, p. 283)
TABLE 1.1  KINDS OF VALIDITY PROPOSED IN 1954

<table>
<thead>
<tr>
<th>Kind of validity</th>
<th>Aim of measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content validity</td>
<td>Determine how an individual would perform at present in a given universe of situations of which the test situation constitutes a sample</td>
</tr>
<tr>
<td>Predictive validity</td>
<td>Predict an individual's future performance</td>
</tr>
<tr>
<td>Concurrent validity</td>
<td>Estimate an individual's present status on some variable external to the test</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Infer the degree to which the individual possesses some trait or quality</td>
</tr>
</tbody>
</table>

defined, but instead could and should be defined implicitly by a network of associations, the nomological network.

The first three kinds of validity were associated with the already known validation procedures. Content validation usually involved expert opinions, and concurrent and predictive validity referred to the correlation with a (simultaneously or later assessed) criterion. Construct validation on the other hand required formulating predictions on the basis of the nomological network and gathering empirical data to confirm these predictions. Thus, the test and the underlying theory were simultaneously validated. Because many predictions could be formulated on the basis of the underlying theory, construct validity could not be expressed in one single coefficient.

Campbell and Fiske (1959) reacted to that simultaneous validation of test and theory. They remarked that, before one could turn to testing the relationships between traits (construct validation), one needed some confidence in the measurement of the trait. Validation by the multitrait-multimethod matrix would yield such a confidence in the measurement. This procedure is aimed at showing that the trait can be differentiated from another trait measured using the same method, and at showing that the trait can be observed using different methods.

1960-1980: TRIPARTITE SCHEME OF VALIDITY

In the following twenty years the discussion about validity and validation became rather quiet and it seemed a period of consolidating and refining the ideas proposed in the fifties. In 1955, Cronbach and Meehl pointed out that predictive validity and concurrent validity might be considered together as criterion-oriented validation procedures. The American Psychological
Chapter 1

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Definition of Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1927</td>
<td>Kelley</td>
<td>Whether a test really measures what it purports to measure</td>
</tr>
<tr>
<td>1937</td>
<td>Bingham</td>
<td>The closeness of agreement between the scores and some other objective measure of that which the test is used to measure</td>
</tr>
<tr>
<td>1954</td>
<td>APA</td>
<td>The degree to which the test is capable of achieving certain aims</td>
</tr>
<tr>
<td>1974</td>
<td>APA</td>
<td>The appropriateness of inferences from test scores or other forms of assessment</td>
</tr>
<tr>
<td>1985</td>
<td>APA</td>
<td>The appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores</td>
</tr>
<tr>
<td>1989</td>
<td>Messick</td>
<td>An integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment</td>
</tr>
</tbody>
</table>

Association (1966) followed this viewpoint and, up until today, validity is discussed using the resulting tripartite scheme of validity (content validity, construct validity and criterion-related validity).

During the sixties and the seventies the definition of validity slowly shifts. Cronbach and Meehl already noted in 1955 that one does not validate a test, but only a principle for making inferences (p.297). It took the American Psychological Association two decades to follow this point of view. In 1974 validity is defined in the Standards as:

"The appropriateness of inferences from test scores or other forms of assessment"

This changed definition marked a move from the aims of testing towards philosophy. The soundness, worth or efficacy of a test was no longer the main issue of validity, the emphasis lay on the reasoning based upon test-scores. According to the Standards and theorists, not only validity itself, but also the validation procedures should become detached from the specific test aims. The idea, that a thorough study of a test should involve information on all three kinds of validity, became emphasized with increased strength over the years.

1980-1990: EXPANSION OF THE CONCEPT OF VALIDITY

In the eighties a renewed interest was taken in validity. In the United States psychological tests were widely used for selection, of army personnel, of
The concept of test-validity

FIGURE 1.1  NUMBER OF PUBLICATIONS IN THE FIELD OF PSYCHOLOGY IN THE 20TH CENTURY AND NUMBER OF PUBLICATION DEALING WITH VALIDITY AS REPORTED IN PSYCHLIT

Number of publications

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1900</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1910</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1920</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1930</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1940</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1950</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1960</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1970</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1980</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1990</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

students, of job applicants, and of jurors, to mention a few selection situations. The widespread use of tests resulted in a public debate challenging the fairness and appropriateness of test use and even resulted in lawsuits (see Cronbach, 1988; Langenfeld & Crocker, 1994; Sokal, 1987). As a consequence, validity theorists started to advocate that test devisers were obliged to guard against potential adverse consequences and should consider the social fairness of test-scores (e.g., Cronbach, 1988; Messick, 1980). In 1989, Messick even includes those consequences into his definition of validity (see Table 1.2).

In fact, the entire field of psychology, and hence psychological testing as well, had experienced a rapid growth. The use of psychological tests (especially achievement tests), the number of articles dealing with test-validity (see Figure 1.1), and the number of different kinds of validity kept growing. Despite the effort of the American Psychological Association in the early fifties to put a halt to the number of different kinds of validity, Van Berkel (1984) categorized in the mid-eighties 77 different kinds of validity.

This rapid increase gave rise to a strong need for evaluating the present status of validity. In the second part of the eighties several elaborate reviews
were published (Anastasi, 1986; Angoff, 1988; Messick, 1989, among others). Within these articles concerns regarding the testing practice were expressed. Theorists feared that test constructors leaned backwards contently once they had covered the three aspects of validity (if at all). Therefore, these articles enunciated that the tripartite scheme of validity should be abandoned. Construct validity had become the all embracing concept. Construct validation was described as a never-ending process, beginning from the outset of test-development.

“What has come to be designated construct validity is actually a comprehensive approach that includes the other recognized validation procedures – and much more besides” (Anastasi, 1986).

BRIDGING THE GAP WITH PRACTICE

During the past century the discussion about validity has moved away from methods of validation to theorizing about the concept of validity. Indeed, nowadays some refer to the discussion, as the philosophy of validity (Moss, 1992). In fact, the validity movement has been one of denominational segregation; nowadays we speak of test users, test constructors and validity theorists or philosophers. The validity-discussion stayed alive, thanks to the tension between these three sections, which is well illustrated by Messick’s (1988, p. 34) remark:

“The difficulty and disagreement alluded to earlier occurs not so much at the theoretical level of validity [...]. Rather, the disagreement is more forcefully, though indirectly, revealed by the persistent disjunction between validity conception and validation practice”.

In the eighties the tone of that discussion had become accusing and venomous. For example, Messick (1988, p. 36) wrote:

“A pessimist might view the current state of testing practice as blatant hypocrisy”

In the early nineties this tone was abandoned and at the theoretical level an integrative movement emerged (e.g., Anastasi, 1993; Ellis & Blustein, 1991a, 1991b; Messick, 1995). Both test-development and even test use are now considered part of validation. Furthermore, the relevance of recent advances in other research areas, i.e. psychometrics and cognitive psychology (see Hambleton & Zaal, 1991), is stressed. In order to discuss all that comes into
play with validation, new categorizations of aspects, issues and criteria of validity and validation, are used (see for a review, Moss, 1992).

Although the tone of voice has become one of acknowledgement and recognition, theory and practice are still far apart. In practice it is rather difficult knowing to which theoretical notion one should comply, because these notions keep shifting. Furthermore, it is difficult to know how to comply, given the comprehensive view of validity. The distance between the testing theory and practice is illustrated by the findings of Qualls and Moss (1996). They examined the extent to which testing practices complied with professional guidelines regarding examination of relevant reliability and validity evidence. They found that only 50% of the instrumentation was supported by evidence of one type or the other.

A potential cause for the discrepancy between theory and practice is caught in the word “comply”. Test validation and test development should comply with the theoretical notions of validity, but a truly cohesive and integrated approach demands equal partnership. One theorist, who actually did build a bridge between theory and practice, is Embretson (1994). She presented a framework, the Cognitive Design System, for utilizing cognitive theory for the development and validation of ability tests. A similar framework for the development and validation of personality and attitude measures is still lacking.

Looking at the history of this scientific discussion about validity and validation we find that the definition of validity has changed. The theoretical notions of both the object of validation and the recommended way of validation shifted. Furthermore, we see that those theoretical notions drift apart from the practice of test validation and test development. Presently, we are facing the challenge of bridging the gap between theory and practice and increasing the cohesion between the various research areas important to validity.