On testing plausible threats to construct validity

Schouwstra, S.J.

Citation for published version (APA):
When, as a student, I was asked to validate a test, I did not foresee that the subject of validity would fascinate and rivet me for years. I thought, I just had to find out whether the test measured what it was supposed to measure. Indeed, this was the first definition of validity. And like so many before me, I discovered things were not that simple. In Chapter 1 the difficulties encountered in validation research in the 20th century and the subsequent changes in the definition of validity are sketched.

Traditionally, the validity question is raised a posteriori. A psychologist sets out with a question or hypothesis concerning an attribute or quality of persons. In order to answer that question or test the hypothesis, this attribute has to be measured. Therefore, a measurement instrument is designed, a psychological test. This test is then administered to subjects and their responses are registered and analyzed. Finally, the results of the analysis are interpreted in terms of the attribute of interest. Then one starts to wonder whether the test score can be interpreted in terms of that attribute and whether the test score adequately reflects the attribute. How trustworthy is the proposed test score interpretation? This question refers to what is nowadays known as construct validity. Construct validity can be defined as an evaluative judgement of the trustworthiness of a test score interpretation.

Construct validation research is aimed at investigating the trustworthiness of a test score interpretation, usually using empirical data. In empirical psychological research three research strategies are used (Mellenbergh, 1980). The first strategy is to support the interpretation with empirical data. In general, the supportive research strategy is a rather weak strategy, because one will always find some support and one does not have criteria to reject the hypothesis (Mellenbergh, 1980). The second research strategy is to test the hypothesis using empirical data. This strategy enables us to reject a hypothesis, but rejecting a hypothesis without formulating an alternative is rather unrewarding. This brings us to the third research strategy, in which alternative hypotheses are formulated. The empirical data are used to test the alternative hypotheses and to select one. Despite the fact that this research strategy overcomes the weakness of the supportive research strategy and is more rewarding than the refutation strategy, it is hardly used.
The first, supportive, research strategy is the most frequent occurring strategy in construct validation research, like in other research areas. The oldest form of empirical support for a test score interpretation is predictive success. At first the evidence focussed at the predictive success of the measurement of one single attribute, but in the fifties this univariate approach was replaced with a multivariate approach (see Chapter 1). Validation researchers now seek support for a network of relationships between the measure of interest and measures of other attributes and criterion-measures. Often, the first data set obtained for validation is explored and used to generate a model of those relationships. Validation research seldom goes beyond this exploratory stage, although it provides only initial support for the trustworthiness of a test score interpretation.

When a relationship between the measure of the attribute and other attributes is found, the test score might still reflect only part of the attribute and the test score might also reflect something more than the attribute of interest alone. In other words, such initial support as results from the supportive research strategy does not exclude the two global threats to validity (Messick, 1995): construct under-representation and construct irrelevant variance. Therefore, after initial support for a test score interpretation, more powerful research strategies should be used to exclude these two threats.

The validation studies presented in Chapter 2 and 3 focus at construct representation. In these studies, the third research strategy of testing alternative hypotheses is employed. In both chapters, alternative models of the structure imbedded in a measure of health locus of control are tested. In Chapter 2 the alternative models are tested in a sample of respondents that differs from the sample used in the initial validation study. In Chapter 3 the alternative models are tested using a different set of items. Therefore, these two studies use a stronger research strategy for testing the structure and support the generalizability of the construct representation across samples of respondents and items.

The validation studies presented in Chapters 4 and 5 deal with the second global threat to construct validity: irrelevant variance. More specifically, these studies focus at potential irrelevant variance caused by the measurement method. In Chapter 4 the response scale of the Photo Anxiety Questionnaire is the subject of validation. An experiment was conducted to test whether or not subjects can discriminate accurately between the photographs used as a response scale. In Chapter 5 the validation research is aimed at refuting that
the responses are influenced by the presentation order of items and the presentation order of photographs forming the response scale.

Sequentially addressing possible treats to construct validity using more powerful research strategies, renders a test score interpretation more trustworthy. But how do we know when to stop? Often, validation stops too early, as research tends to stop after the first initial support. In contrast, we can continue testing against alternative hypotheses into absurdity, because one can formulate, a posteriori, infinitive alternative ways in which the test score might have been influenced. Each step and each point in the process from theoretical hypothesis to test score interpretation might influence the test score and invalidate the resulting interpretation. The various points in this process are, retrospectively, scrutinized for potential threats.

A disadvantage of this a posteriori approach is that we tend to formulate the most obvious or noticeable alternatives, rather than the most plausible alternatives. We are inclined the pay close attention to the final phase of the process, the analyses, and the sample. The first phase, from hypothesis concerning the attribute to the measurement of the attribute, tends to escape our attention. The rationales underlying the test score interpretation are seldom questioned. What is more, the links between the various steps in the process from hypothesis to interpretation are not addressed.

In Chapter 6 a framework for construct validation research is presented, the Deductive Design. The Deductive Design aims at increasing the trustworthiness of a test score interpretation by increasing the cohesion and consistency between the two lines of scientific evidence, rationales and empirical evidence. The central issues of construct validity, construct representation and irrelevant variance, are simultaneously addressed from the outset of test development. In Chapter 7 the Deductive Design is applied to the development and validation of a self-efficacy scale.

All the studies presented in this thesis are aimed at increasing the trustworthiness of a test score interpretation by addressing the two global threats with more powerful research strategies. In the four chapters, following the historical overview in Chapter 1, this aim is tackled a posteriori. In the last two chapters this problem is approached a priori, from the outset of test development. All chapters have been written as separate papers. Hence, some overlap between the successive chapters can occur.