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

Citation for published version (APA):
When the work for my thesis drew to an end, family and friends often posed me the question: “What problem did you solve?” A question they, upon seeing my hesitant and uncomfortable expression, quickly rephrased as: “What problem did you clarify?” Well, I guess I did neither. The main problem with validity and validation, if you can call it a problem, is that it never ends. As Anastasi (1986) put it: “Test-validity is a living thing; it is not dead and embalmed when the test is released” (p. 4).

The never-ending character of validity and validation is not attributable to the sublimity of the issue, to be addressed adequately by a true genius only. No, it is because of the same reason why so many of us take an interest in psychology: the diversity and change of people and the diversity and change of their thoughts. We will have to validate new and old instruments alike as long as our measurements use language or require an active and conscious involvement of the people whose qualities we are measuring.

We have to continue validating, because both components of construct validity (construct representation and irrelevant variance) can and will change in the course of time. Any change of thought in one of the theoretical fields underlying validation requires a renewed evaluation of the validity of existing test-score interpretations. New psychological constructs and theories will be developed, and the existing ones can be redefined or abandoned. Such changes imply a redefinition of the construct underlying the construct representation. New psychometric models and statistical techniques may change the way our test-scores are most appropriately modeled and analyzed. New methods and theories concerning test-development and validation may be developed. All these theoretical developments and changes of thought will have an impact upon the construct representation.

The sources of irrelevant variance are apt to change as well in the course of time. The language we use to phrase our questions and items can become hopelessly old-fashioned within 10 years and, consequently, introduce group-type bias. Furthermore, the knowledge, experiences and social norms of people change. For example, when I started as a researcher in psychology, differences in experience with the computer would still greatly influence the results of self-administered computer tests. The majority of students had to be instructed, first, on the use of the mouse, an instruction that in seven years time has
become almost redundant. Similarly, cultural norms of what is appropriate or decent to say can change very rapidly and differ widely. For example, asking dichotomous (Yes/No) questions would be rather inappropriate in China, due to unforgivable impoliteness of saying "No". Such a question-format would yield an assessment of nothing.

The construct representation and sources of irrelevant variance can change, likewise the definition of validity may show further changes. The discussion on the concept of validity is not free of the influence of language and culture, as I discovered. Messick's new definition of validity (1989) has received much endorsement in the circle of those concerned with validity theory and educational testing in the United States, although in some publications we get a glimpse of a dispute held in private (Cronbach, 1988, p. 6; Messick, 1988, p. 34). Many Dutch methodologists, on the other hand, were perplexed at his proposal to include values, potential consequences, and fairness into the definition of validity and reacted with great reserve. This reserved reaction is defendable on scientific grounds (e.g. Lees-Haley, 1996), but the perplexity with which the Dutch reacted had also cultural roots. A public debate about test use, as occurred in the United States, has not occurred in the Netherlands, because test use is not so widespread and extensive. Furthermore, social pressure in the form of lawsuits has not occurred.

Part of the perplexity, however, definitely lies in the particular meaning of the words "validity" and "valid" in Dutch. These words have a formal meaning in Dutch and are hardly used in everyday language. The Dutch use is limited to valid reasoning or laws on the one hand and to a person's physical fitness or ability to work on the other hand. The words "efficacy", or "worth" are not associated with the word "validity" in Dutch, neither does the word "valid" has an equivalent meaning like the English "just". The expansion of the definition of validity, as proposed by Messick, is consequently more perplexing to a Dutch person than to an English-speaking person.

As the reader may have concluded by now, I definitely learned to steer away from answering a question. Let me rephrase the initial question myself: "What contribution does my work offer?" The research presented in this thesis is for a large part dedicated to validation procedures advocated, and acknowledged as important, but seldom published or employed. The most stringent example is the replication study (Chapter 2 and 3). Seldom, such studies find their way into journals, despite the fact that the importance of replication is recognized. Rosnow and Rosenthal (1989) wondered whether
there were sociological grounds for the monomaniacal preoccupation with the results of a single study. They hypothesized that this preoccupation had to do with the reward system of science (psychological science). Support for this hypothesis can be found in the publication manual of the American Psychological Association (1994).

The first criterion for deciding whether a research merits publication, mentioned in the manual, is whether the research question is significant, and whether the work is original and important. Given the scarcity of replication studies one can view replication studies as very original, but seldom do we associate replication studies with originality. Replication studies are often viewed as merely a reinforcement of an argument and subsequently as little important. Consequently, that first criterion of the publication manual does not favor replication studies.

The reward system is probably not the single cause for only paying lip service to the importance of replication studies. The importance of replication is a well-learned lesson in scientific education, but I have begun to suspect that the rationale for its importance and the very nature of replication escapes some psychologists. I found reviewers to wonder why it would be important to confirm a factor structure and to criticize that the study followed the earlier study too narrow. As Rosnow and Rosenthal (1989) pointed out, our discipline would be farther ahead if it adopted a more cumulative view of science. Replication studies are requisite for such a cumulative approach, and, furthermore, they give insurance for the robustness and generality of earlier findings.

The objections raised against replication studies do not apply to the studies presented in Chapters 4 and 5. However, such studies into irrelevant variance are even more seldom than the (replication) studies into construct representation. The tradition to raise the construct validity question only after a test has been developed might underlie this scarcity. We risk accepting a test score interpretation as trustworthy, before the most plausible threats have been formulated and excluded, when we sequentially address possible threats in separate studies. Furthermore, this a posteriori approach leads to an overemphasis upon empirical validation strategies. The rationales underlying a test score interpretation escape our attention.

The Deductive Design presented in Chapter 6 is aimed at insuring that the most plausible threats to construct validity are addressed simultaneously and at increasing the cohesion between both scientific arguments, rationales and
empirical evidence. The application of the Deductive Design, presented in Chapter 7, demonstrates the value of the framework for rendering an evaluative judgement of the trustworthiness of a test score interpretation. Furthermore, and maybe more importantly, this study demonstrates that the Deductive Design enables us to reach the conclusion that a test score interpretation is untrustworthy. A conclusion that probably was not reached had the problem of construct validity been approached sequentially and a posteriori.

Validation studies in general give insurance for the robustness and generality of specific test score interpretations, but also contribute to both scientific debate and cohesion, the very soil of scientific progress. Validation research takes place at the cross-roads between the various fields within psychology, therefore such research can uncover discrepancies, assumptions, and pitfalls within the different theoretical notions and in doing so give rise to a scientific discussion. As was noted in Chapter 1, the scientific conduct is often marked by consensus within the separate fields, and by disjunction between the fields (see Chapter 1). It is the disjunction that leads to scientific discussion and hence to progress.

Validation is not and cannot be limited to the mere application and use of theories originating from various fields. Test-development and validation require the translation and integration of ideas and notions from those various fields. For that reason, test research contributes to the cohesion within psychology. A cohesion that is strongly needed, among the multitude and abundance of ideas, models and mini-theories, as De Groot (1981) among others pointed out. Hopefully, the research presented in Chapter 6 and 7 will enhance cohesion between testing practice and testing theory that drifted apart during the 20th century. Furthermore, developments in validity and validation theory were increasingly based upon achievement and ability testing only. I hope my work is instrumental for re-involving the field from which the concept of construct validity originated (personality and attitude assessment) into the discussion on validity.