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

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 6.

THE DEDUCTIVE DESIGN: A DEFINITION-DRIVEN APPROACH TO TEST-DEVELOPMENT AND CONSTRUCT VALIDATION

Abstract. - The unified concept of validity and the never-ending process of validation are difficult to handle when dealing with construct validation in practice. The Deductive Design proposed here presents a framework for construct validation, when dealing with constructs for which the concept of construct validity was originally introduced, such as personality and attitudes. Construct validity is an evaluative judgement of the trustworthiness of a test score interpretation. Two kinds of scientific arguments can support such an interpretation: rationales and empirical evidence. Within the Deductive Design both rationales and empirical evidence center on the test score interpretation as derived from the underlying theoretical notions. Furthermore, construct validation requires that two issues be addressed: construct representation and irrelevant variance. Within the Deductive Design both issues of construct validity are addressed from the outset, prior to test development. Because construct validity is an evaluative judgement, the Deductive Design incorporates also an integrated evaluation of both scientific arguments (rationales and empirical evidence) covering both issues, construct representation and irrelevant variance. This integrated evaluation delineates subsequent lines of construct validation research and provides a sound basis for research on the validity of other score based inferences and test utility.

INTRODUCTION

At the beginning of this century the problem of validity was defined as the problem whether a test really measures what it purports to measure (Kelley, 1927). In the years to follow, three different kinds of validity were distinguished, each associated with a different validation procedure: content validity, predictive validity and concurrent validity. These three kinds of validity were strongly linked to a specific aim of testing, that is, performance assessment, prediction, or classification. The concept of construct validation was introduced because the associated validation procedures were difficult to apply to tests that were primarily designed to measure theoretical constructs. In the first description of construct validity, it was almost described as an escape route:
"Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured." (Cronbach & Meehl, 1955).

According to Cronbach and Meehl construct validation was involved whenever a test was to be interpreted as a measure of some attribute or quality which was not "operationally defined". These attributes could be defined implicitly by a network of associations, the nomological network. In construct validation, both hypothesis and predictions should be formulated on the basis of the nomological network and should, subsequently, be tested empirically.

In Cronbach and Meehl's first article on construct validity, several important ideas were formulated, albeit carefully, that eventually gave rise to the present view of validity and construct validity (see for an overview, Angoff, 1988). In the first place, they stated that it is not the test, that is validated, but a principle for making inferences. This idea became widely accepted in the years to follow. Validity is now defined in the Standards for Educational and Psychological Testing (1985) as the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores.

Secondly, Cronbach and Meehl stated that, in construct validation, we seek to justify how one is to defend a proposed interpretation of a test (p.284). In this view many types of evidence are relevant to construct validity, including content validity and criterion-related validity. The present view is that these different kinds of validity are not only relevant to construct validity, but that construct validity actually embraces them. Messick (1995) remarks that construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores. These many lines of evidence cannot be integrated into one single coefficient; they can only be integrated into a judgement. Construct validity can therefore be defined as an evaluative judgment of the trustworthiness of a test score interpretation in terms of a construct.

This present view yields three difficulties when dealing with construct validation in practice. First of all, in principle all evidence gathered in the process of developing and using a test has bearing on the interpretation of test scores.

---

1 Validity theorists often refer to the more global definition of Messick (1989). He defines (p. 13) validity as an integrated evaluative judgment 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.
scores (ea. Anastasi, 1986, Messick, 1989). Furthermore, as long as a test continues to be used new evidence is gathered. Therefore, construct validation never ends. Third, test score interpretation starts with the construct definition and consequently this definition is the first line of evidence (Anastasi, 1986). This implies that the process of construct validation starts prior to test development, hence the integrated evaluative judgement has to include the rationales or evidence pertaining to the process of test development. And finally, the problem that gave rise to the formulation of construct validity still exists. We are still facing the difficulty of obtaining content- and criterion-related evidence for theoretical constructs that cannot be “operationally defined”.

Given these difficulties, it is not surprising that practice falls short of the theoretical demands. The abundance of theoretical changes, recommendations, and the multitude of potential pitfalls and validation methods have made it difficult to see the wood for the trees. In order to clarify construct validation Messick (1995) distinguishes six aspects. These aspects afford a means of checking whether the central issues of validity are addressed irrespective of the specific construct, aim or measurement mode. Although these distinctions clarify our general view, they do not prescribe or describe the specifics. Stated in other words, they do not provide a clear-cut approach for a construct validation procedure in which these aspects are addressed and linked.

Embretson (1994) does provide a clear approach, but it is aimed at ability tests only. In her Cognitive Design System the emphasis is placed on linking test-development and construct validation with cognitive theories. For abilities, observable phenomena can be postulated to have a very strong covariation with the construct of interest, namely solving specific problems correctly or incorrectly. In other words, abilities can be measured using so-called maximum performance tasks (Cronbach, 1990)\(^2\), which a third person can evaluate as having been solved correctly or incorrectly. The Cognitive Design System is not directly applicable, and not intended to be so, to the development and validation of instruments for other kinds of constructs.

A similar systematic theory-driven approach for development and construct validation is needed for non-ability tests; tests intended for those bothersome theoretical constructs for which the concept of construct validation was proposed in the first place. These theoretical constructs, such as attitudes

\(^2\) I thank G. J. Mellenbergh for this suggestion.
and personality, are measured by so-called typical response tasks (Cronbach, 1990), because they refer to phenomena or attributes within individuals that cannot be observed by a third person, and for which no other observable, strongly related phenomena can be found.

THE DEDUCTIVE DESIGN FOR CONSTRUCT VALIDATION

As already stated, construct validity can be defined as an evaluative judgement of the trustworthiness of a test score interpretation in terms of a construct. According to De Groot (1981), judgement formation occurs through exchange, criticism and argumentative discussion within the (scientific) Forum. Hence, a judgement of trustworthiness of a test score interpretation occurs through a scientific argumentative discussion\(^3\). An argumentative discussion entails that scientific arguments have to be presented pertaining to the interpretation. Such scientific arguments should, by preference, focus at plausible rival hypotheses (see Campbell, 1986; Cronbach, 1988) or threats.

Two global threats to a test score interpretation can be distinguished (Messick, 1995, among others). When interpreting a test score in terms of a theoretically defined construct, the test score is regarded as a reflection of the construct of interest. The first threat, construct underrepresentation, occurs when the test score reflects only part of the construct as defined. The second threat, irrelevant variance, occurs when the test score reflects more than the construct of interest, which means that some of the test score variance is irrelevant to the construct of interest. So, the scientific arguments should support the interpretation that the test score reflects the whole of the construct as defined, and, second, reflects nothing else.

This interpretation can be supported by two kinds of scientific arguments: rationales and empirical evidence. Within Embretson's Cognitive Design System both the rationales and evidence centered on the cognitive processes required for solving the maximum performance task. Likewise, the framework presented here, focuses at the link between the interpretation and the construct definition. The Deductive Design is a framework for validation of a test-score, designed to measure a theoretical construct, which is obtained with a typical response task. The interpretation of such a test score is based in the first place upon the theory concerning the construct of interest.

\(^3\) Several validation theorists (Cronbach, 1988; Kane, 1992) take such an argumentative perspective of validity.
Within the Deductive Design, therefore, the rationales (the first scientific argument) consist of specifying how the interpretation is derived from the underlying theoretical notions. This requires, first, formulating what the construct of interest is and what not. But, the interpretation is based also upon the operationalization and modeling of the construct of interest. Therefore, the way in which the construct definition (and nothing else) is translated into the test content has to be specified next. Third, we have to model how the typical responses to the test content reflect the construct of interest and nothing else. So, in fact, the derivation of the test-score interpretation is threefold: formulation, translation and modeling (see Figure 6.1).

The second scientific argument consists of presenting evidence that supports the derived interpretation. This entails presenting evidence that the test-score reflects the whole of the construct and evidence that the test-score does not reflect something else.

When the first line of empirical evidence is presented construct validation has not ended, because construct validity is not a list of arguments, but an evaluative judgement. Consequently, a framework for construct validation should move beyond summing up the initial arguments supporting an interpretation or refuting an alternative interpretation. The Deductive Design encompasses, therefore, an initial evaluation of construct validity that indicates future lines of construct validation research. This initial evaluation entails evaluating and integrating the eight components of the argumentation displayed in Figure 6.1.

60
The Deductive Design is a framework for obtaining a judgement of the trustworthiness of a test score interpretation, as derived from the construct definition, by evaluating an argumentation. The argumentation involves specifying how the interpretation is derived from the identification, translation and modeling of the construct of interest (the rationales), and, second, presenting evidence supporting the derived interpretation. Both arguments should cover the two central issues denoted by the threats to the derived interpretation, that is complete construct representation and the absence of irrelevant variance.

FORMULATION

Construct representation

The construct of interest has to be identified and defined. The construct of interest can be defined in two ways. Kerlinger (1986) defines a construct as an abstraction formed by generalization from similar phenomena or attributes that is systematically defined to be used in scientific theory. So, a construct is defined explicitly by the group of phenomena or attributes the construct refers to. Furthermore, the construct is implicitly defined by the other constructs and phenomena of the scientific theory, of which the construct is part (Cronbach & Meehl, 1955).

Generally we use a very global theory and apply this theory to a more specific research domain. Consequently, the researcher should first specify the group of attributes the construct refers to and next how this group of attributes, the construct domain, is narrowed down to fit the specific research domain (Hox, 1997). For example, the construct “attitude” is defined (see Guttman, 1982) and then this definition is narrowed down to “political attitude”. The researcher should also make clear how the construct is assumed to exert its influence on other phenomena. This entails determining the structure or essential aspects of the construct. The essential aspects are those aspects that can cause a difference in the influence the construct exerts. For example, the extremity of attitudes (from negative to positive), the accessibility of an attitude in memory (Fazio, 1989) and the strength of an attitude (Krosnick and Abelson, 1992) can exert influence on behavior. The exact specification of the group of phenomena or attributes a construct refers to and the structure of that group will be called the explicit construct definition.
Chapter 6

The formulation of such an explicit construct definition is difficult when dealing with vaguely described constructs, which is often the case. Clarification of a construct (see Hox, 1997) can be obtained by dimensions/indicator analysis (Fiske, 1971; Lazarsfeld, 1958, 1972), semantic analysis (Sartori, 1984), signifying concept analysis (De Groot & Medendorp, 1986), or facet theory (Guttman, in Levy, 1994). We should note that most of these techniques are not limited to the exact formulation of a construct, but also include the operationalization.

After formulation of the explicit construct definition, the construct should also be defined in the second way. The implicit construct definition has to be formulated. The construct is part of a scientific theory, which constitutes of a set of interrelated constructs, definitions and propositions that present a systematic view of phenomena with the purpose of explaining and predicting the phenomena (Kerlinger, 1986). The set of constructs, through which or in relationship with which the construct of interest exerts its influence, and the phenomena, one seeks to explain or predict, delineate the boundaries of the construct of interest. These boundaries provide an implicit definition of the construct. For example, in the theory of reasoned action (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975) attitudes and subjective norms influence intentions. Intentions on their turn influence behavior. Within that theory subjective norms and intentions thus implicitly define attitudes.

And, as was the case with the explicit construct definition, the researcher should specify how the implicit definition is narrowed down to fit the specific research domain.

Irrelevant variance

After identifying the construct of interest, we have to identify the phenomena or constructs of disinterest (see Figure 6.1). Because obviously all else is of disinterest, we have to identify what is of particular disinterest given the construct and task for measuring the construct. The range of interest is limited by the theory concerning the construct of interest and by the particular research question. A construct is defined for a particular purpose, and,

---

4 This example illustrates that not all constructs and relationships from the nomological network are included in the implicit construct definition. The implicit construct definition contains only those constructs and relationships through which the construct of interest exerts its influence. As a consequence, the implicit construct definition is narrower than the nomological network.
furthermore, a construct is proposed in order to explain, predict or select within a particular population, setting or time. In other words, the formulation of the *purpose and conditions* of the research identifies the limits of interest given the construct.

Before turning to the actual translation of the construct definition into test content, we need to identify what is of particular disinterest given the task for measuring the construct. A task has to be selected and the skills and attributes, that are required for responding to the selected task, have to be identified. The choice of the task is strongly guided by the kind of construct the researcher intends to measure. The Deductive Design is intended for constructs that are measured by tasks requiring a typical response. Typical response tasks are tasks with which we seek to determine a person's typical performance, i.e., what that person is likely to do in a given situation (Cronbach, 1990). Usually these typical response tasks rely on self-report, for example a person is asked directly to indicate whether or not he/she intends to vote. However, other tasks are possible. Dovidio and Fazio (1992), for example, have developed procedures to obtain an indirect assessment of attitudes.

The selected task determines the skills and attributes required for generating a response. Some of these skills (or all of them) do not reflect the construct of interest. For example, the response process with verbally presented tasks requires language skills. These requisite skills can influence the test scores and cause irrelevant variance. This kind of irrelevant variance will be referred to as *concept-related irrelevant variance*, to denote that the irrelevant variance is caused by skills or attributes that are not part of the scientific theory concerning the construct of interest and hence are not part of the implicit construct definition.

Summarizing, the formulation should provide an answer to four questions:

1) To what group of attributes does the construct refer and how does the construct exert its influence (the explicit construct definition)?

2) Which phenomena does the construct effect in interplay with what other constructs (the implicit construct definition)?

3) Why is the construct defined and for whom, where and when (the purpose and the conditions)?

4) What are the skills and abilities required for generating a response to the selected task?
TRANSLATION

After the formulation, the construct definition has to be translated into observable variables appearing to be a suitable representation of the construct (Hox, 1997). This translation should be based upon all four formulated aspects: the explicit and implicit construct definition, the purpose and the required skills.

Irrelevant variance

In order to prevent irrelevant variance due to the skills required for response generation (concept-related irrelevant variance), guidelines for the translation have to be formulated. If, for example, one uses written statements as a task, the reading skills of the respondents might influence the test scores. In that case, the guidelines could be that all items have to be equally difficult in terms of item wording, and that the item wording has to be easy enough for the target population to understand. More generally phrased, the level of ability (or skill) required for generating a response should be minimal and the different items of the test should require the same level of ability.

Secondly, the method of test administration has to be determined. The specified purpose and conditions partially determine the administration method (mode, procedure, and item order). Usually, the administration method is also based upon personal preferences and pragmatic reasons, such as costs. But the administration method can also be aimed at preventing alternative non-intended reactions. For example, a written questionnaire instead of an interview might be preferred, in order to prevent interviewer effects. The testing situation can be viewed as a social event (Rosenthal and Rosnow, 1969). As a consequence, the respondent might not only react to the manipulated item attractiveness, but also to the testing situation. This may yield, for example, responses that are influenced by social desirability or by test anxiety. In other words, the social character of the testing situation may yield irrelevant variance also. This kind of irrelevant variance is due to concepts that are not required for generating a response, and that are, consequently, difficult to anticipate.

Construct representation

In accordance with the guidelines the construct definition has to be translated into the item content. The researcher should seek to vary only
content referring to the construct of interest and should try to exclude any content that refers to the constructs that appear in the implicit definition. So, on grounds of the implicit construct definition, boundaries of the item content have to be formulated. For example, when the attitudes are implicitly defined by social norms and intentions (i.e. in the theory of reasoned action), the items measuring an attitude should not be phrased in terms of "I will". Such a phrasing refers to a construct (i.e. intention), that appears in the implicit definition.

A systematic translation of the explicit construct definition can be obtained by means of a facet design. Examples of such a facet translation can be found in Stouthard, Mellenbergh and Hoogstraten (1993) for dental anxiety and with Hox (1986) for subjective wellbeing. The attractiveness or difficulty of the items should reflect different levels of the various aspects of the construct.

Furthermore, the responses to these items have to be registered and scaled. Registration requires determining the number and labels of response categories. The labels should be consistent with the type of response (discrete or continuous) invoked by the item content. A simple strategy to detect a mismatch is to reformulate the item as a question and to use the response labels to formulate an answer. The labels match the item content, when it is possible to formulate a meaningful answer to the question. To illustrate, in the next item a mismatch between item content and response labels occurs.

Circle the number that describes the degree to which you feel happy

1 definitely do not feel 2 cannot decide 3 slightly feel 4 definitely feel

The item can be reformulated as the question: How happy do you feel? But only the third response label yields an appropriate answer. The first and fourth labels, for example, yield answers to the question: How certain are you of your feeling of happiness?

The scaling of the responses should be consistent with the registration of those responses. Once again, one should be suspicious of mismatches. For example, the responses to the item mentioned above were scaled at interval level. An interval level implies that the four categories can be ranked and that

---

5 Personal communication with G. J. Mellenbergh, 1999
6 The survey interview with typical response tasks which is the focus of this article is viewed as a special case of conversation (e.g. Bingham & Moore, 1934; Bradburn, 1992).
7 This item is cited from Russell and Carroll (1999), who use the item as an example of ambiguous polarity.
the differences between the subsequent categories are equal, which is very doubtful in the given example.

Summarizing, like the formulation, the translation involves four steps. The identified requisite concepts, the implicit construct definition, the explicit construct definition and the purpose and conditions are translated into guidelines, boundaries, items and method of administration respectively. The entire process is directed at translating the construct definition into observable variables in such a way, that all potential respondents can give a response as intended. This entails that the attractiveness of items has to reflect the various aspects and levels of the construct as defined in the explicit construct definition. Furthermore, the reflection or influence of the other constructs from the implicit construct definition and of (required and non-required) concepts associated with the task and administration method should be minimized or excluded.

MODELING

Construct representation

After the translation, we should turn to the final step of formulating the rationales. As many validity theorists have emphasized, the researcher should specify exactly how the test score reflects the construct of interest in a model (see among others Loevinger, 1957, on structural fidelity; Messick, 1995, on the structural aspect of validity and Embretson, 1983, on modeling). The translation of the explicit construct definition can be represented in a measurement model. The choice of a specific measurement model requires that the aggregation level of the responses is specified. A model for sumscores or a model for item responses might be used (e.g. Mellenbergh, 1998). When a model for sumscores is used, the way in which the responses are aggregated should be specified, based on the theoretical specifications of the construct. The test score should adequately reflect the different aspects of the construct.

If an item response model is used a choice is required among the various possible models. The choice is usually only partially delimited by the formulation and translation (for example the choice between a model for dichotomous responses or polytomous responses) and is often a matter of individual preference. As Hox and Mellenbergh (1990) point out, more attention should be paid to the link between operationalization and the choice
of a particular item response model. An example of linking the operationalization and the choice of a model is the step approach (Van Engelenburg, 1997). In this approach, a choice among the various models for polytomous ordinal item responses is linked to different task characteristics.

But not only the formulation and translation of the explicit construct should be represented in a model. The same goes for the implicit definition and the additional constructs. The implicit construct definition can be represented in a structural model (i.e. Bollen, 1989) in which the nature, strength, and direction of the relationships of the construct of interest with the other constructs and phenomena are specified.

*Irrelevant variance*

Likewise, a structural model for the additional concepts has to be formulated. In the formulation and translation, several required and non-required concepts were identified whose influence on the test score should have been minimized. Therefore, the second structural model should represent the absence of association of the variable (construct) of interest with the variables of those additional concepts.

The administration procedure and the identified purpose and conditions can impose restraints upon the method of data-collection. For example, a particular sample or method of sampling might be required or a particular research design. Usually, the exact method of data-collection is determined when turning to empirical evidence.

To conclude, all three rationales (identification, translation and modeling) cover the two central issues, construct representation and irrelevant variance, denoted by the threats to a test score interpretation, as displayed in

Figure 6.2. The issue of complete construct representation is addressed by formulating the group and structure of attributes a construct refers to (the explicit construct definition). Furthermore, the phenomena and constructs that are influenced by the construct of interest are identified (the implicit construct definition). The explicit construct definition is translated into item content and into a response scale. Next, the explicit construct definition is represented in a measurement model. The implicit construct definition is translated into boundaries of item content and, next, represented using a structural model.


Chapter 6

Figure 6.2  The Scientific Argumentation within Deductive Design

<table>
<thead>
<tr>
<th>Scientific arguments</th>
<th>Rationales</th>
<th>Construct Representation</th>
<th>Irrelevant Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical evidence</td>
<td>Implicit</td>
<td>Explicit</td>
</tr>
<tr>
<td></td>
<td>Empirical evidence</td>
<td>Fit Structural Model</td>
<td>Item Relevancy</td>
</tr>
<tr>
<td></td>
<td>Empirical evidence</td>
<td>Comparison</td>
<td>existing Instrument</td>
</tr>
<tr>
<td></td>
<td>Empirical evidence</td>
<td>Model</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Modeling</td>
<td>Structural Model</td>
<td>Measurement</td>
</tr>
<tr>
<td></td>
<td>Modeling</td>
<td>Structural Model</td>
<td>Measurement</td>
</tr>
<tr>
<td></td>
<td>Translation</td>
<td>Boundaries</td>
<td>Items &amp; Scale</td>
</tr>
<tr>
<td></td>
<td>Translation</td>
<td>Implicit</td>
<td>Explicit</td>
</tr>
<tr>
<td></td>
<td>Formulation</td>
<td>Implicit</td>
<td>Explicit</td>
</tr>
<tr>
<td></td>
<td>Formulation</td>
<td>Definition</td>
<td>Definition</td>
</tr>
</tbody>
</table>

Issues

Concept-related

Method-related

Comparison other Method

Data-collection

Method of Administration

Purpose & Conditions

68
The Deductive Design

The issue of irrelevant variance is addressed by identifying the skills required for generating a response and by identifying the research conditions and purpose. These requisite skills are translated into guidelines for item wording and represented in a structural model. The purpose and conditions, on their turn, are translated into an administration method and data collection method.

PRESENTATION OF EMPIRICAL EVIDENCE

The first argument for the scientific discussion (the rationales), to render a judgement of the trustworthiness of a test score interpretation, is now presented. Next, the second argument has to be presented: the empirical evidence. On the basis of the specification of the rationales empirical data are collected to test the interpretation. The sample, design, and procedure of the data-collection should be consistent with the formulated purpose and conditions, as well as with the specified administration procedure. As for the measures to be taken, the construct of interest should be measured using the measurement instrument and using an existing instrument, if available, for comparative purposes (the multi-method comparison, Campbell & Fiske, 1959). Furthermore, the constructs and phenomena of the implicit construct definition and the identified requisite and non-requisite concepts should be measured (the multi-trait comparison, Campbell & Fiske, 1959). Finally, the collected data are analyzed by fitting the specified models. The techniques chosen to analyze the data should be consistent with these specified models.

Like the rationales, the evidence also covers the two central issues: evidence that the test score reflects the whole of the construct (construct representation) and evidence that the test score does not reflect anything else (irrelevant variance).

Construct representation

The first line of empirical support consists of evidence that the test score reflects the whole of the construct as defined explicitly and implicitly. A global assessment of the explicit construct representation is obtained by comparing the new instrument with an existing instrument for the construct of interest, the so-called multi-method comparison. The implicit construct representation, on the other hand, is assessed globally by fitting the structural model. Once again the new measurement instrument for the construct of interest can be
compared with an existing instrument by fitting the structural model using both instruments.

This global evidence has to be supplemented with specific evidence, that is, evidence at an item level. Confirming the explicit construct representation involves confirming the hypothesized structure and item attractiveness, since the items are specified to reflect different aspects of the construct and to reflect different levels of these aspects. The explicit construct definition can be confirmed by fitting the specified measurement model.

Also, specific evidence regarding the implicit construct representation has to be obtained. The (essential) aspects of the construct are hypothesized to cause a difference in the influence the construct exerts on observable phenomena. Therefore, the hypothesized relevancy of the aspects of the construct should be confirmed by assessing the relationship of each aspect with the phenomena to be predicted.

Irrelevant variance

The second line of evidence is aimed at refuting that the test score reflects something other than the construct of interest. The first alternative hypothesis is that the test score reflects or is influenced by other concepts, which is called here concept-related irrelevant variance (see Figure 6.2). The concepts required for responding should not cause variation in the test scores. Neither should the non-required concepts, associated with the social character of the testing situation, cause variation in test scores. In order to refute the alternative interpretation of concept-related irrelevant variance, the test scores have to be compared with the test scores reflecting those other concepts.

But evidence that other concepts do not cause substantial differences in test scores does not imply that the intended response is actually given. Even when all requirements for excluding concept-related irrelevant variance are met, the intended response might not be given. The many potentially confounding response effects can be roughly grouped into two additional ways in which the actual response can deviate from the intended response.

The intent is that the respondents react to the attractiveness of the item, as manipulated by the researcher. The actual response deviates from the intended response, when the respondent reacts to the non-manipulated attractiveness of the item. At least three sources of non-manipulated attractiveness exist. The first source is the attractiveness of the response categories. When respondents react to the attractiveness of the response categories, response styles such as
yes-saying or satisficing (Krosnick, 1991) are the result. Second, the attractiveness of an item might be influenced by the attractiveness of the previous item, which results in item order effects. And third, the attractiveness of the item content or of the response categories might have changed due to repeated administrations (see Bray, Maxwell, & Howard, 1984; Koele & Hoogstraten, 1988; Schwartz & Sprangers, 1999 on the subject of response shift and Schmitt, 1982; Schmitt, Pulakos, & Lieblein, 1984, on beta- and gamma-change).

The second way in which the actual response can deviate from the intended response occurs when the response is relative rather than absolute. The response might for example be given in reference to a group. This yields relative item bias (Borsboom, Mellenbergh & Van Heerden, 1999), also called group-type bias (Oort, 1994, 1996). The response might also be given in reference to earlier answers, which results in an increase in the consistency between the responses (the so-called Socratic effect, see among others, O'Malley & Thistlewaite, 1980).

These forms of irrelevant variance, method-related irrelevant variance, can be assessed globally by comparing several methods of test administration. For example, by comparing two administrations of the measurement instrument with different item orders. At item-level both concept-related irrelevant variance and method-related irrelevant variance (for an overview of all types of bias see Oort, 1994, 1996) can be assessed using various item bias detection techniques (Holland & Wainer, 1993; Millsap & Everson, 1993), e.g. restricted factor analysis (Oort, 1992).

Summarizing, evidence regarding the explicit and implicit construct representation and evidence regarding concept-related and method-related irrelevant variance has to be presented (see Figure 6.2). The data are collected in concordance with the formulation and translation. Following, these data are analyzed globally and specifically using the specified models and item bias techniques.

INITIAL EVALUATION

Once the two scientific arguments, rationales and empirical evidence, have been presented, an evaluation has to be presented that subsumes these arguments and that indicates subsequent lines of validation research. The whole argumentation addresses two basic issues of an interpretation: (explicit
and implicit) construct representation and (concept-related and method-related) irrelevant variance. The researcher sets out with specifying the rationales, the derivation of the interpretation, which comprises the theoretical formulation, the translation into the test and the modeling of the test score. And next the global and specific empirical evidence is presented (see Figure 6.1). So, evaluating the whole argumentation requires evaluating several components and the links between the components. This evaluation consists of three aspects. The first aspect deals with the question whether all components of the argumentation support the interpretation, the ambit of the argumentation. The second aspect focuses at evaluating how strongly these components support the interpretation, the strength of the argumentation. And third, the generalizability of the components has to be evaluated.

**Ambit**

The ambit of the argumentation can vary due to two causes. First of all, one of the issues of the argumentation, for example concept-related irrelevant variance, might be completely absent. Complete absence means that neither the rationales nor evidence supporting the issue was provided. But also a link might be missing, which results in a missing-component in only one of the two arguments. For example, a multi-trait comparison might be performed, although the potentially confounding concepts were not identified in the specification. Of course, a complete argumentation enhances the trustworthiness of an interpretation. The Deductive Design for construct validation aims at preventing such an absence of components.

The second cause for variation in ambit is the possibility of finding only partial support for the test-score interpretation. For example, the researcher may have found support for the construct representation in terms of the explicit construct definition, but not in terms of the implicit construct definition. Or, no support may have been found for the exclusion of the concepts required for generating the responses. The item analyses show whether the lack of complete support might be caused by some of the items. Then the ascertainment of partial support should result in item selection or item rewriting.

As is often emphasized, item selection and rewriting should not be carried out solely on psychometric grounds. Whereas the evidence of the item-analyses yields the psychometric grounds, the specification yields the theoretical rationales for selection and rewriting. Of course, the rules and rationales for
selection and rewriting have to be clearly specified and have to be unambiguous in order to insure replicability. With clear and unambiguous rules an arbitrary other researcher should retain the same set of items given the same data set. Clarity furthermore insures that the appropriateness of selection and rewriting can be tested in separate lines of research.

**Strength**

How strongly confirmation or lack of confirmation effects the trustworthiness of the proposed interpretation depends on the precision of the argumentation. The more precise the researcher has been about the construct of interest and about *how* the constructs and variables are related and vary, the more difficult it is to obtain confirmatory evidence. For example, the presence of a relationship between two variables is more easily confirmed than the presence of a linear relationship with a particular strength and direction. In the latter case the relationship between the variables is specified to be restricted to a particular kind of relationship. Thus, the more precise (restrictive) the argumentation is the more supportive evidence adds to the trustworthiness. In case of complete, but weak confirmatory evidence subsequent research should be directed at increasing the precision of the argumentation by adding restrictions.

Another matter of importance is that we should be circumspect of missing links between the arguments, as was the case with the ambit. In articles on validation of particular tests, the occurrence of three kinds of non-linked restrictions can be noted. The first are contradictory restrictions. An example of such a contradiction is the use of an analysis technique for linear relationships when in fact the relationship between two constructs is defined as non-linear.

The second kind is a restriction that could have been deduced but was not. An example of such an absent restriction is the use of an exploratory factor analysis even though a specific structure is defined and implemented in the test. If a specific structure is implemented, this structure should be translated into the measurement model and a confirmatory factor analysis should be used. The Deductive Design is aimed at preventing these first two kinds of non-linked restrictions, which could result in lack of complete confirmation.

The third kind of non-linked restrictions are restrictions that are not present in the formulation but added later on in the argumentation. We can differentiate between added restrictions that are prerequisite and those that are not. An example of an added restriction which is not prerequisite occurs when
a highly restrictive measurement model is chosen (e.g. the Rasch-model, 1960), whereas no theoretical rationales support those restrictions and a far less restrictive model (e.g. Mokken, 1970, 1997) could have been used.

But, in order to obtain and analyze evidence, choices and restrictions have to be made even when explicit theoretical rationales lack. This lack of theoretical rationales can occur at various levels. Mellenbergh (1980) pointed out that when a psychological theory is investigated using empirical data, theories at other levels, that is, psychometric theories and data-analytical theories, play an important role. Generally, the psychological theory about the construct of interest is not and probably cannot be specific enough to deduce the prerequisite restrictions and choices imposed at the other theoretical levels. As already pointed out, the method of administration and the choice for a particular measurement model are difficult to deduce from the identification and translation. Furthermore, the assumptions underlying particular measurement models might not have been established yet (see Van de Brink & Mellenbergh, 1999). Also, the analysis techniques used are based on specific assumptions regarding, for example, the distribution of the variables. Often, these assumptions cannot be linked with the psychological theory.

This kind of an added non-linked restriction can also cause lack of complete confirmation. The evident course of subsequent action is to choose an alternative analysis technique, measurement model or administration method. Of course, haphazardly choosing alternatives is not recommended. When theoretical rationales fall short, the choice can be based on simulation studies on robustness against the violation of assumptions underlying particular measures, or underlying the measurement model and analysis techniques (e.g. Van Engelenburg & Schouwstra, 1997).

**Generality**

The final characteristic of the argumentation is its generality, which determines the generalizability of the test score interpretation. The generality of the argumentation is limited in two ways. First, the generality of the whole argumentation can not exceed the generality of the final argument: the presented evidence. The generality of the evidence is almost always smaller than the generality of the rationales. Replication studies are needed to establish whether the findings are as general as defined. Usually, we are aware of the limits imposed by the method of gathering evidence (e.g. the particular sample of respondents), even though this awareness seldom results in replication
The Deductive Design

studies. But also a particular method of test administration, a particular set of items and a particular measurement model are chosen. These choices influence the findings as well and are grounds for replication (De Leeuw, Mellenbergh, & Hox, 1996). Strictly speaking, every choice made after the formulation that was not deduced, or that was not based upon random sampling, requires a replication.

Second, the generality of the argumentation is limited to the implicit construct definition and to the purpose and conditions mentioned in the specification. In other words, the test score cannot be interpreted solely in terms of the explicit construct definition, but can only be interpreted within the context of the psychological theory and the particular research question. If the construct of interest is placed within another psychological theory, the boundaries of the construct and processes through which the construct exerts its influence are defined differently, and therefore the interpretation is by definition different. Therefore, great reserve should be employed with respect to formulating theoretical changes. All too often lack of confirmation is attributed to the psychological theory. Above several potential causes for lack of confirmation are outlined which should be investigated before turning to formulating new psychological theories or psychological models.

Applicability of the Deductive Design

The Deductive design is intended to provide in the need for a systematic theory-driven approach to the development and construct validation of tests for theoretical constructs, such as attitudes and personality. In this framework, construct validation is addressed before the instrument is developed. Therefore, the Deductive Design is characterized by the a priori approach to construct validation. It is essential in this approach that not only construct representation is dealt with from the outset of test development, but that also the irrelevant variance, that is plausible given the task and measurement method, is dealt with from the outset. Furthermore, in order to obtain a judgement of the trustworthiness of a test score interpretation (a judgement of construct validity), the various sources of support have to be evaluated and integrated. Therefore, the second important characteristic of the Deductive Design is that it incorporates such an integrated evaluation of rationales and empirical evidence (the scientific argumentation).
In contrast to Messick’s aspects of construct validity (1995) and, for example, Kane’s argumentative approach (1992) to validity, the applicability of the Deductive Design is limited to a specific kind of construct, which is measured using a specific kind of task. Just like Embretson’s Cognitive Design System (1994) is only intended for ability tests with maximum performance tasks, the Deductive Design proposed here is only intended for non-ability tests with typical response tasks. But the applicability of the framework is further limited due to the aim to start construct validation with the construct definition. Oosterveld and Vorst (1996; Oosterveld, 1996) classified the various test development methods according to the nature of the a priori information about the construct. The development can be based on empirical data (the inductive method), on informal knowledge of either experts or potential respondents (the intuitive method) or on formal knowledge (the deductive method). The applicability of the Deductive Design is limited to the deductive test-development method, because formal knowledge, the construct definition, is taken as the basis for test-development.

This latter characteristic of the Deductive Design is another difference with Embretson’s Cognitive Design System. In the Cognitive Design System, it is not the construct definition, but the response process required for solving the task that is taken as the basis for test-development. The cognitive process used to solve a task is reflecting the construct of interest and thus provides a definition of the construct. Within the Cognitive Design System the test-development is aimed at ensuring that variance in the test score reflects differences in this response process. In contrast, the focus of the Deductive Design is at excluding test score variance due to these response processes, since these processes do not reflect the construct of interest.

Construct validation within the Deductive Design is aimed at providing arguments for the derived test score interpretation that the whole of the construct is represented and that no irrelevant variance is present. Emphasis is placed upon the specification of the rationales underlying the interpretation, which entails formulating the construct definition and sources of irrelevant variance, translating these into the test content, and representing them in models. The empirical evidence to be presented should support this derived interpretation. The link between specification and evidence ensures detection or prevention of mismatches and provides an understanding of empirical findings. Finally, construct validity can be assessed by evaluating the whole argumentation, which delineates subsequent lines of construct validation.
research and thus provides a sound basis for research into the validity of other score based inferences and test utility. As such, the Deductive Design provides a framework for obtaining an initial judgement of the trustworthiness of a test score interpretation as derived from the construct definition.