

TYPES OF VALIDITY

There's an awful lot of confusion in the methodological literature that stems from the wide variety of labels that are used to describe the validity of measures. Firstly, it is not very wise to limit our scope only to the validity of measures. We really want to talk about the validity of any operationalization, that is, any time we translate a concept or construct into a functioning and operating reality (**the operationalization**), we need to be concerned about how well we did the translation. This issue is as relevant when we are talking about treatments or programs, as it is when we are talking about measures. Secondly, we could certainly use the term construct validity to refer to the general case of translating any construct into an operationalization. Let us use all of the other validity terms to reflect different ways you can demonstrate different aspects of construct validity.

With all that in mind, here is a list of the validity types that are typically mentioned in texts and research papers when talking about the quality of measurement:

Construct validity

Translation validity

Face validity
Content validity

Criterion-related validity

Predictive validity
Concurrent validity
Convergent validity
Discriminant validity

Translation validity is just a seemingly sensible name to summarize what both face and content validity are getting at. All of the other labels are commonly known.

Let's see if we can make some sense out of this list. First, as mentioned above, let us use the term construct validity to be the overarching category.

Construct validity is the approximate truth of the conclusion that your operationalization accurately reflects its construct. All of the other terms address this general issue in different ways. Second, we make a distinction between two broad types: translation validity and criterion-related validity. That's because these correspond to the two major ways we can assure/assess the validity of an operationalization.

In **translation validity**, we focus on whether the operationalization is a good reflection of the construct. This approach is definitional in nature - it assumes we have a good detailed definition of the construct and that we can check the operationalization against it.

In **criterion-related validity**, we examine whether the operationalization behaves the way it should given your theory of the construct. This is a more relational approach to construct validity. It assumes that our operationalization should function in predictable ways in relation to other operationalizations based upon our theory of the construct. Let's go through the specific validity types.

In **face validity**, we look at the operationalization and see whether "on its face" it seems like a good translation of the construct. This is probably the weakest way to try to demonstrate construct validity. For instance, we might look at a measure of math ability, read through the questions, and decide that yep, it seems like this is a good measure of math ability (i.e., the label "math ability" seems appropriate for this measure). Or, we might observe a teenage pregnancy prevention program and conclude that, "Yep, this is indeed a teenage pregnancy prevention program." Of course, if this were all we do to assess face validity, it would clearly be weak evidence because it is essentially a subjective judgment call. (Note that just because it is weak evidence doesn't mean that it is wrong. We need to rely on our subjective judgment throughout the research process. It's just that this form of judgement won't be very convincing to others.) We can improve the quality of face validity assessment considerably by making it more systematic. For instance, if we were trying to assess the face validity of a math ability measure, it would be more convincing if we sent the test to a carefully selected sample of experts on math ability testing and they all reported back with the judgement that our measure appears to be a good measure of math ability.

In **content validity**, we essentially check the operationalization against the relevant content domain for the construct. This approach assumes that we have a good detailed description of the content domain, something that's not always true. For instance, we might lay out all of the criteria that should be met in a program that claims to be a "teenage pregnancy prevention program." We would probably include in this domain specification the definition of the target group, criteria for deciding whether the program is

preventive in nature (as opposed to treatment-oriented), and lots of criteria that spell out the content that should be included like basic information on pregnancy, the use of abstinence, birth control methods, and so on. Then, armed with these criteria, we could use them as a type of checklist when examining our program. Only programs that meet the criteria can legitimately be defined as "teenage pregnancy prevention programs." This all sounds fairly straightforward, and for many operationalizations it will be. But for other constructs (e.g., self-esteem, intelligence), it will not be easy to decide on the criteria that constitute the content domain.

In **criteria-related validity**, we check the performance of our operationalization against some criterion. How is this different from content validity? In content validity, the criteria are the construct definition itself -- it is a direct comparison. In criterion-related validity, we usually make a prediction about how the operationalization will perform based on our theory of the construct. The differences among the different criterion-related validity types are in the criteria they use as the standard for judgment.

In **predictive validity**, we assess the operationalization's ability to predict something it should theoretically be able to predict. For instance, we might theorize that a measure of math ability should be able to predict how well a person will do in an engineering-based profession. We could give our measure to experienced engineers and see if there is a high correlation between scores on the measure and their salaries as engineers. A high correlation would provide evidence for predictive validity - it would show that our measure can correctly predict something that we theoretically think it should be able to predict.

In **concurrent validity**, we assess the operationalization's ability to distinguish between groups that it should theoretically be able to distinguish between. For example, if we come up with a way of assessing manic-depression, our measure should be able to distinguish between people who are diagnosed manic-depression and those diagnosed paranoid schizophrenic. If we want to assess the concurrent validity of a new measure of empowerment, we might give the measure to both migrant farm workers and to the farm owners, theorizing that our measure should show that the farm owners are higher in empowerment. As in any discriminating test, the results are more powerful if you are able to show that you can discriminate between two groups that are very similar.