

Steps in Creating a New Psychometric Measure

When you want to create a new psychometric measurement tool (e.g., health status, homophobia, patient satisfaction), there are several specific steps. Below is a common approach.

Preliminary investigation:

Always begin with some qualitative work. Unstructured interviews with people relevant to the measure. If you want to develop a new measure for how breast cancer affects quality of life, the most important group to talk to is people with different stages of breast cancer, however, interviews with experience health care personnel may also be informative. These interviews can be done individually or in focus groups.

Initial Item Generation:

Based upon the preliminary qualitative investigations and previous research or measurement tools, individual measurement items are generated that covers the spectrum of what you want to measure. Some duplication should be retained, since you do not yet know the best way to phrase the questions, however, the questions should be screened for clarity, comparability with each other, and not have so much duplication so as to breed responder revolt.

Initial Pilot Testing:

The initial pilot testing should be done in a small number (maybe 5-25, depending upon your budget) of subjects who will give you feedback on ambiguities, offensive questions, difficult or time-consuming portions of the instrument).

Preliminary Study:

After revising the instrument based upon pilot testing, you decide whether more pilot testing is needed, or whether you proceed with a preliminary study. The study population should be representative of the population that you eventually want to use the instrument on. The study should be large enough to do at least preliminary reliability testing. Factor analysis is often used to shed light on how many different concepts or domains are being measured. Once it is determined which measures are most correlated with each other (and therefore may belong in the same scale), you look at the items (checking for face or content validity) and name the scale. If evaluator and evaluatee are the same (e.g., self-reported health status), than test-retest and internal consistency (usually measured by Cronbach's Alpha) of the scale should be measured. If the evaluator is external to what is measured (e.g., satisfaction with doctor, medical student evaluation), than inter-rater reliability (usually measured by Kappas or correlation coefficients) should also be evaluated.

Validation:

Validation analyses are much more difficult and messy than reliability testing. The validity and limits of a new measurement tool usually takes many studies to sort out. See the Sage Publication on Reliability and Validity for a general discussion.

Item reduction:

A high level of reliability (Alpha) is only needed for individual judgments (e.g., a measure of who should have surgery). If aggregate assessments are the only goal (e.g., do surgery patients do better than medically treated patients) than reliabilities greater than 0.8 are often wasteful and fewer questions or evaluators can be used. The random measurement error can be accounted for in the statistical analyses. You are usually still better off if your measure has a reliability of better than 0.7, but lower reliabilities do not necessarily invalidate the results.

Using Factor Analysis to Help Guide Scale Development

Factor Analysis = A technique for examining the interrelationship of a set of variables. All variables are treated equally (there are no DepVars and IndVars) and the object of the analysis is to create *factors* that help explain the multivariable associations between the original variables.

How is factor analysis used? There is controversy regarding how factor analysis should be used, but it is most commonly used when developing psychometric scales. Most researchers feel that factor analysis should mainly be used to *help* test and confirm that the items (such as survey questions) measure the conceptual domain that you set out to measure. It should be used in conjunction with theory and common-sense and should not be used to try to avoid theory and common-sense. This is probably best demonstrated by example.

Example 1: You asked 10 questions about attitudes about homosexuality trying to measure homophobia. Factor analysis can be used to see if all 10 items are associated, as hypothesized, with a single domain, *homophobia*, or whether there is evidence that there are two or more different scales or whether some items do not measure the intended domain. In one study, factor analysis suggested that there are two related, but separable, domains of homophobia that were named “social homophobia” (being uncomfortable around someone who is gay) and “moral homophobia” (feeling that being gay is immoral). It also found that one item (attitudes about “whether homosexuals were mistreated by society”) had a high negative correlation with homophobia in the U.S. but was uncorrelated with homophobia in France. Psychometric scales try to measure psychological and socially constructs, therefore, what items go with which items can vary dramatically between cultures and subcultures.

Example 2: You ask residents to rate their attending using 20 questions. These 20 questions try to distinguish 1) how competent the attending is (fund of knowledge and clinical skills), 2) how good a teacher they are, and 3) the extent to which they teach using medical evidence vs opinion. Factor analysis can be used to determine whether the questions used appear to weight most heavily with the intended scales and the degree to which there is evidence for the 3 intended scales. In several studies, it has been found that residents do not tend to distinguish between the competence, knowledge and teaching ability of attendings. This does not mean that these separate domains do not exist, only that it appears that learners perceptions in these settings do not appear to distinguish between these concepts when evaluating attendings.