# 10 - 9. Psychometry of rating scales

# 9. Psychometry of rating scales:

© SPMM Course 9. Psychometry of rating scales: When developing measurement scales, we are concerned about two important properties. Can we use this scale to measure the actual phenomenon we want to measure? Can this scale provide consistent results when it is used? A highly valid scale will measure what it is supposed to measure – the truth. A highly reliable scale will provide consistent results. Reliability refers to the replicable nature of research studies / tools. Note that high reliability does not guarantee scientific validity but guarantees consistency.  Reliability can be assessed by test-retest correlation by administering an instrument twice to the same population. The time difference between test and retest must be long enough to avoid practice effect, but short enough so the underlying state (e.g. depression) does not change very much: 2 to 14 days range is often used in psychiatry.  Cronbach's alpha measures the internal consistency of a test by correlating each item with the total score and averaging the correlation coefficients. It can take values between negative infinity and 1 as a maximum; but only positive values make sense. Arbitrary cut-off of 0.70 is used commonly to call the evaluated test to be internally consistent.  The split-half reliability refers to splitting a scale into two parts and examining the correlation.  Interrater reliability is measured using two or more raters rating the same population using the same scale.  The intraclass correlation coefficient is used for continuous variables; it is nothing but the proportion of total variance of the measurement that reflects true between subject variability. It ranges between 0 (unreliable) and 1 (perfect reliability). ICC can be measured by either relative agreement or absolute agreement; the relative ICC is always higher than the absolute ICC. ICC of 0.6 is considered fair while 0.8 is very good and 0.9 as excellent, arbitrarily. ANOVA intraclass coefficient is used for quantitative data with more than 2 raters/groups.  For nominal data that has more than two categories, a kappa or weighted kappa can be used. (More details are given below) Validity of an instrument is the extent to which an instrument measures what it proposes to measure.  Face validity refers to a subjective measure of deciding whether the test measures the construct of interest on its face value.e.g., Hamilton depression scale clearly has a face value in measuring depression; but not for measuring obsessions.

© SPMM Course  Construct validity measures whether a test really measures the (theoretical) construct of interest or something else. One way of classifying the construct validity is considering unified construct validity. Here construct validity is taken to consist of both content validity and criterion validity (referred as unified construct validity).  Content validity refers to whether the contents i.e. each individual subscales, items or elements of the test are in line with the general objectives or specifications the test was originally designed to measure. It looks for a good coverage of all domains thought to be related to the measured condition. This often cannot be statistically tested, but experts are called for comments on this aspect of validity.  Criterion validity refers to the performance against an external criterion such as another instrument (concurrent) or future diagnostic possibility (predictive).  Concurrent validity refers to the ability of a test to distinguish between subjects who differ concurrently in other measures (using other instruments). e.g., those who score high on a scale of insomnia may score high on a scale of fatigue ratings too.  Predictive validity refers to the ability of a test to predict future group differences according to current group differences in score. e.g., high aggression score in childhood and high criminal incidents in adult life. (On a similar note, Incremental validity refers to the ability of a measure to predict or explain variance over and above other measures)

Another way of considering the construct validity is by classifying it to convergent, discriminant and experimental/interventional validity:  Convergent validity refers to agreement between instruments that measure same construct e.g. between BDI and HAMD for depression. This agreement can be tested in contrasted groups i.e. depressed and non-depressed, both groups showing a high correlation between the two scales.  Discriminant validity refers to the degree of disagreement between two scales measuring different constructs. e.g., to say that HAMD measures some construct (depression) different from that measured by Hamilton Anxiety scale (anxiety) poor correlation must be demonstrated between HAMD and HAS  Experimental validity: This refers to the sensitivity to change. An instrument must show the difference in results when an intervention is carried out to modify the measured domain.

© SPMM Course Note: Factorial validity is a form of construct validity established via factor analysis of items in a scale. Precision and accuracy Precision is the degree to which a calculated central value (e.g. mean) varies with repeated sampling. The narrow the variation, the precise the value is. Random errors lead to imprecision. Factors reducing precision includes 1. Having wider the limits of the interval 2. Expecting higher confidence interval (e.g. 99.7% versus 95%). Accuracy refers to the correctness of the mean value – i.e. how close is it to the true population value. Precision is comparable to reliability while accuracy is comparable to validity. Bias in a study compromises validity / accuracy. VALIDITY QUESTION IT ANSWERS Face Does this scale appear to be fit for the purpose of measuring the variable of interest? Content Does this scale appear to include all the important domains of the measured attribute? Criterion Is the scale consistent with what we already know (concurrent) and what we expect (predictive)? Convergent Does this new scale associate with a different scale that measures a similar construct? Discriminant Does the new scale disagree with scales that measure unrelated constructs?

---

Revision #1
Created 2026-01-04 20:05:41 UTC by Omar Ayman
Updated 2026-01-04 20:05:41 UTC by Omar Ayman