

Notes on Reliability and Validity of the Delaware Student Testing Program

Reliability: the extent to which test scores are consistent, or, the degree to which the test scores are dependable or relatively free from random errors of measurement.

1. Reliability of .85 or above is expected on instruments such as the DSTP.
2. Reliability on reading and mathematics tests is greater than .9 on both tests, all forms.
3. Reliability above .9 helps ensure accurate classifications of students and greatly reduces the possibility of miscalculations when cut points are established.

Validity: the extent to which a test does the job for which it is intended.

1. ***Content Validity: the extent to which the content of the test represents a balanced and adequate sampling of the outcomes about which inferences are to be made.***
 - a. Primarily refers to the match between what is taught and what is tested. However, in the standards movement this often refers to what should be being taught (i.e., a set of standards) and what is tested.
 - b. Each item on the DSTP is validated for a content match to the standards prior to being field tested by teacher/development groups and external measurement professionals. Following field testing each item is further validated using data and student responses to ensure that the match is accurate. In addition, studies are run annually using various rater groups to calibrate the standards match and the strength of that match, and the results are used to inform the development process.

Note: with any good testing program validity is a process and not a product. Staff members who work directly on the DSTP instruments understand this and spend much of their time refining and fine tuning, resulting in an instrument that is thoroughly defensible from the perspective of content validity in a standards-based environment.

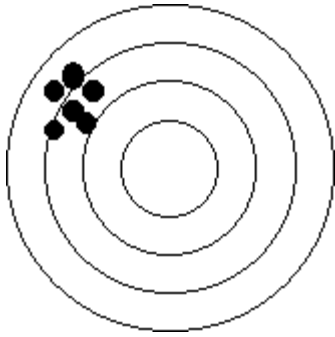
2. ***Criterion-related validity: the extent to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) some criterion measure.***
 - a. i.e., the accuracy to which a test is indicative of performance on a future criterion measure (predictive validity).
 - b. i.e., the agreement between the test being validated and some collection of data (concurrent validity).

Note: tests are not validated for either of these until after they have been given. Studies of this type will be conducted as data become available.

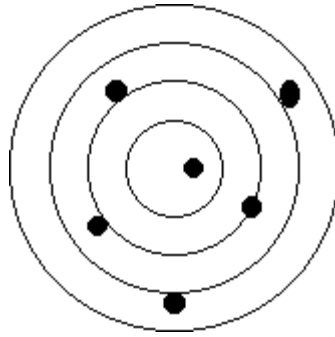
3. ***Construct Validity: the extent to which a test measures some relatively abstract or psychological trait or construct.***
 - a. Generally this kind of validity is determined by determining the relationship between test scores and pertinent external data.

Note: tests are not validated for construct validity until after they have been given. Studies of this type will be conducted as data become available.

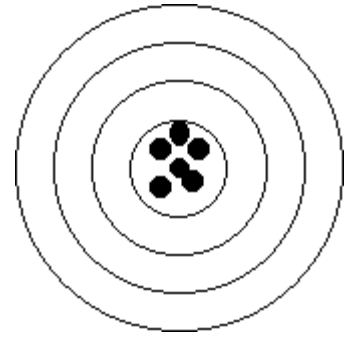
A common diagram used to communicate the relationship between reliability and validity is as follows:



Reliable but not valid



Neither valid nor reliable



Valid and reliable