Chapter 2
Validity

Introduction

Before I discuss the importance of validity and how it relates to those who practice clinical neuropsychology, a definition is in order. Validity is an umbrella term with many tributaries, all of which have slightly different meanings and methods for determining their legitimacy (Nunnally & Bernstein, 1994). However, the linchpin of all validation studies in neuropsychology is the ability to operationally define and uncover unique dimensions of cognitive architecture that serve to describe, explain, and predict. Real-world experiences (including affect, behavior, and cognition) are sampled by neuropsychological tests. Thus, someone who experiences word-finding difficulties in conversation is likely to demonstrate such difficulties on tests of naming (e.g., the Boston Naming Test) (Kaplan, Goodglass, & Weintrab, 1983). This test, among all tests in neuropsychology, is predicated on the assumption that occurrences outside of the examination room are captured on neuropsychological testing, and that these measures represent bona fide surrogates for "life experiences" and brain function. This is validity.

Hypothetical Constructs

A hypothetical construct is something believed to exist despite an inability to measure or observe this phenomenon directly. As in psychodynamic thought where we are faced with unobservable, latent constructs such as the id, ego, and superego, in neuropsychology, we must reconcile with what we are unable to assess unalloyed phenomena—including anxiety and executive cognition. Despite neuropsychology's appearance of perhaps more "scientific" and quantifiable pursuits (e.g., akin to the physical sciences), we are still left in the uncanny position of having to develop measures that serve as proxies for a plethora of neuropsychological constructs such as memory, attention, as well as spatial, configural, and executive cognition.

D.A. Maroof, *Statistical Methods in Neuropsychology: Common Procedures Made Comprehensible*, DOI 10.1007/978-1-4614-3417-7_2,

Since neuropsychologists are compelled into indirect observation, they must backpedal and determine whether the methods cultivated to quantify and measure neuropsychological constructs are indeed measuring what they are claimed to measure.

Fundamentally, neuropsychologists want to ensure that the constructs measured are the sole cause of variability in test performance: depression (or lack thereof) causes scores on a scale for depression to fluctuate, as an example. And, if someone is having problems with attention, then this difficulty presumably should be exerting a causal influence on tests requiring attention and concentration. Thus, test are themselves the operational definitions for the constructed they measure.

But, What Is Validity?

As in any other discipline, the overarching definition of validity within neuropsychology has to do with authenticity—do our tests, scales, indices, measures, etc. measure what we say they measure? Does a list learning test measure memory? Does an index that we have developed to measure depression, actually measure depression, and if so, does it measure different aspects of depression, such as affect (feelings of sadness), behavior (social withdrawal), and cognition (negative thoughts)?

Even if we have established that our measures adequately embody these hypothetical constructs, assessing what we want them to, can they reliably distinguish among disorders? These are important empirical questions that we must try to answer in order to ensure neuropsychology as a worthy science. These are questions that can be addressed with validation studies.

This book (or chapter) does not cover reliability, which is the consistency in one's measurement. However, validity presupposes reliability. On the other hand, a measure can be reliable and yet still entirely invalid. For example, suppose I were to say that hopping on one foot is a reliable measure of intelligence. This may be reliable (if someone can hop on his or her foot for 1 min, they will be or likely to be able to do the same task sometime in the near future). However, it is very unlikely that hopping on one foot is a valid measure of intelligence (in fact, if someone hops on one foot thinking that it is measuring intelligence, then, I guess we are measuring something else (e.g., gullibility?)).

What Are the Types of Validity?

There are several aspects of validity, including content, criterion, construct, incremental, and ecological validity. Validity is a process. It is equally important to remember that validity is a function of the context in which a test is used; it is not entirely tied or attributed to the makeup of a test. It takes two to Tango. Thus, using a test of rapid motor responding in someone with hemiplegia is inappropriate. In this case, it is not the test that is invalid, but the context, which is unsuitable.

As I mentioned, establishing the validity of neuropsychological measures is a process. Items to be included in an instrument considered to reflect the hypothetical construct must be selected (content validity). This can be conceived somewhat as a form of face validity—do the questions or items we have devised seem to measure what we want to measure? However, content validity generally requires more scientific verification. Following decisions on the items for inclusion, the next step may be to determine whether these items truly measure what they are supposed to be measuring (construct validity). This involves examining two important lines of evidence. First, is our construct associated with other measures that assess the same construct (convergent validity)? Second, is our construct unrelated to other measures that should have no relation to our construct (divergent validity)? This can be a somewhat difficult endeavor in neuropsychology because, as you are well aware, many measures in neuropsychology are correlated with one another. Another crucial step in the validation process is to ascertain whether test scores predict conceptually related outcomes. For example, does a scale that measures impulsivity and poor decision making correctly classify those who are incarcerated or involved in other illegal activity?

Two other important method, incremental and ecological validity, minister to clinical neuropsychology. Incremental validity examines how well the addition of certain tests aids in explaining the variation in the dependent variable. For example, does including motivation in an analysis have significant weight in predicting academic achievement over and above intelligence and/or traditional cognitive testing?

Ecological validity on the other hand examines whether neuropsychological tests and constructs relate to real-world functioning. Does a test of executive cognition really reflect someone's ability to reason and problem solve in everyday life (e.g., would a neuropsychological test indicate whether an individual would know what to do should his or her car breakdown?)? Tests that mimic daily functioning (e.g., writing checks or finding phone numbers in a simulated phonebook) have been developed in an effort to offer more face valid assessments with the hopes of capturing how a person actually handles such matters outside of testing. This is an important enterprise for neuropsychologists as neuropsychological testing is often solicited to speculate or render fail-safe decisions on a person's ability to function in a variety of capacities (e.g., live on his or her own versus some form of congregate housing).

Classify or Predict? What's the Big Deal?

Although often used as synonyms, the terms "classify" and "predict" should not be used interchangeably. One is not necessarily predicting something when using data (both test scores and outcome measures)—such as in a regression analysis—that were collected simultaneously. So, if patients with Alzheimer's disease and healthy persons were evaluated on a neuropsychological battery, and then it is determined which tests help discriminate the two groups, the researcher is not really predicting

group membership per se. Instead, he or she is classifying or merely distinguishing between two groups. There is no temporal component to this study, and therefore, this is more of a classification procedure.

The term "predict" should be used, for instances in which the outcome/criterion (such as disease) develops subsequent to measurement of a particular phenomenon. In this sense, prediction means that a score or someone's performance on a particular measure augurs for the person's status (outcome), much like the genetic mutation for Huntington's disease predicts that a person will develop HD—though, not when the person might develop manifest HD (the onset of symptoms). However, when sufficient control is exerted over a number of factors, this bolsters the researcher's confidence that particular tests or findings portend a specific outcome. Thus, knowledge of one piece of information (e.g., test performance) foretells the advent of a particular outcome. It can then be concluded that test performance predicts an outcome with a reasonable degree of certainty.

As another example, patients with Alzheimer disease perform worse on many measures than would normal controls. However, if a cohort of elderly individuals, all healthy, were to complete a neuropsychological battery, and then were followed prospectively, the researcher could establish who develops Alzheimer disease over a certain period of time. The researcher can then determine whether any of the tests the participants completed at baseline predict their *current* status. Here, we can say that a test (or tests) score predicted disease outcome. Note that this is not the same as saying a score *caused* the outcome to occur. Rather, the scores at baseline were likely the cause of a percolating, yet unapparent, inconspicuous illness.

Now let us turn to exploring validity in a bit more detail. While there are other types of validity (e.g., internal, external) we focus on five I mentioned above: content, criterion, construct, incremental, and ecological.

Content Validity

Content validity has to do with how well an instrument is able to capture and encompass all the conceptual aspects of a construct. The researcher must ensure that the items selected for his instrument are an adequate sample from some theoretical content domain. In other words, to assess anxiety, all elements of this construct (physiological, behavioral, and affective) must be included in a measure. However, most researchers in neuropsychology find that the majority of constructs are overly broad; they recognize that to develop one scale or measure to assess every component of a construct would be equally difficult and would likely dilute the significance and meaning of that very construct. This is why more specific aspects of functioning are conceptualized (e.g., set-shifting), rather than tests that measure expansive constructs such as executive functioning proper.

Let us use an example of depression to describe the development of a scale's content validity. The question remains: Is the instrument capturing the affective, behavioral, and cognitive components of depression? There are two methods for which to judge content validity, subjective and objective.

The subjective process would involve asking "experts" within the domain of interest (e.g., depression) to review the instrument's items and decide whether it does indeed cover all aspects of this construct.

An objective measure might be the use of factor analysis to verify the domains that the measure is tapping into. So, after items for the scale have been constructed, and it has been administered to a sample, a factor analysis should reveal the three domains that the instrument was expected to encompass: affective, behavioral, and cognitive dimensions. This is content validity because the measure *fits* with the concept (and content) of depression.

Criterion Validity

Criterion validity has to do with how well an instrument predicts or is associated with an observed indicator of a given concept or criterion (Bryant, 2000). If the measure is genuinely measuring a particular construct, like memory, then, its classification and predictive ability should be well established. An instrument must have an empirical association with some conceptually related criterion.

If a newly developed measure of memory is in fact measuring memory, it should differentiate among persons with amnestic mild cognitive impairment (those with memory deficits only) and those who are cognitively intact. Similarly, if on a measure of daily living skills and independence, one's performance correlates well with his or her real-life adaptive functioning, this demonstrates criterion validity. If a person who is dependent in everyday life, but breezes through the test, then this test may lack criterion validity.

A measure can lack criterion validity for several reasons. First, the content of items (content validity) might be insufficient. Second, the items may not reflect the complexity of real life. For example, people often have much structure in performing tasks when undergoing neuropsychological assessment, whereas they often have to recall and initiate many tasks on their own in real life. One might say that this measure lacks ecological validity—how well a person's performance on neuropsychological testing mirrors or represents how he performs outside of the testing room (cf. Chaytor & Schmitter-Edgecombe, 2003). There are methods for assessing criterion validity, all of which have to do with the temporal ordering of measurement. These approaches include retrospective (postdictive), concurrent, and predictive/prospective validity.

Retrospective/Postdictive Validity

Retrospective validity is fraught with several methodological flaws. The problem with this type of validity is that a construct—a behavior for example—is measured in hindsight, after it has actually occurred. In neuropsychology, a researcher might

be interested in looking at the effects of maternal alcohol abuse during pregnancy and their offspring's performance on cognitive testing at 18 years of age. Thus, tests of the teenagers' current cognitive functioning are possible, but the researcher must rely on available archival data and self-report for additional information (i.e., while their mothers were pregnant). This information might be inaccessible, as it may include acquiring information with regard to how much their mother was drinking at pregnancy, how often, during which trimester (s), and all other salient factors considered might be relevant both during the pregnancy of the mother as well as throughout the course of the participant's (offspring's) development.

Another example would be to compare patients' intellectual function as adults with the number of school absences as a child. The neuropsychologist could assess a group's intellectual function, and then obtain school records from their respective elementary, middle, and high schools. A hypothesis might be that, limiting the study to high school graduates, and controlling for age or significant life-altering factors in school (health problems or family issues), the number of absences is inversely related to adult IQ. That is, the higher the IQ, the less likely the person missed school.

Generally, correlational and regression analyses are appropriate for establishing a measure's postdictive validity. If the variables of interest are highly correlated, or one has significant predictive power, then this supports a measure's postdictive validity.

Unfortunately, there are a number of methodological flaws with postdictive validity. One problem, as you can already imagine, is that memories for past information are subject to massive gaps and distortions. In some cases, archival data (e.g., medical charts) might help reconcile some of these problems.

Concurrent Validity

In this case, the test score and the outcome variable are measured concurrently. For example, one can examine the association (correlational analysis) between self-report of memory difficulties and performance on a test of memory. High correlations would help to establish concurrent validity. One caveat is that if patients were to rate their memory impairment following their actual test performance, the relationships might be spurious. In other words, patients who are depressed often have negative self-appraisals and assume that they did poorly on memory tests despite occasional "normal" performance. On the other hand, patients who perform poorly on memory testing may be well aware that they did not do well on testing, and they may then rate their perception of their memory performance accordingly. Therefore, it is preferred that the test score and outcome measure are obtained from different sources (e.g., a patient and his or her spouse).

Concurrent validity would be more useful with two modality-congruent tests (e.g., two memory tests) rather than mixing and matching the methods of assessment (performance and observer report). The correlations among measures of similar methodologies are likely to be more reliable and valid.

Predictive Validity

As the name suggests, predictive validity establishes how well a particular variable predicts a criterion variable. In this case, the temporal relationship between the test score (which is assessed first) and the outcome measure (assessed last) is essential in order for the researcher to aver that one predicts another. Again, it is not the same to say that a score that precedes a particular outcome caused the outcome. In fact, as Anastasi (1950) argued over 60 years ago, a psychological test is simply a device for determining within a brief period what would eventually emerge with time. For example, we could say that performance on an intelligence test at age 12 predicts whether this person will pursue graduate training. However, the outcome measure (whether or not this person is to pursue graduate training) will reveal itself in due time. Thus, neuropsychological tests are useful in that they provide such predictions of important outcomes, be they academic or medical, well in advance of such protracted observations (Anastasi, 1950).

Using informant ratings of patients' behavior as a predictor of patients developing frontotemporal dementia later in life would be an example of establishing a test's predictive validity. The test score—the informant's report—is culled prior to the criterion (diagnosis of a disease). The method of analysis appropriate for this example would be logistic regression, where the outcome measure is dichotomous (a person does or does not develop the disease). The allure of logistic regression is that it provides both the classification accuracy of the model (how many people who were rated as having behavior difficulties are actually diagnosed with frontotemporal dementia), as well as odds ratios. Odds ratios allow for the examination of an increase or decrease in odds of developing the disease based on informant ratings for every unit increase on the behavior rating scale.

The criterion variable need not be categorical. For example, one can use a scale of postpartum depression to examine levels (gradations) of depression years later (of course, with appropriate covariates in the model). This would involve a multiple regression analysis.

The use of regression models (either linear or logistic) in establishing predictive validity also allows the researcher to explore the contributions of other factors, as well as control for other relevant and possibly confounding factors.

Construct Validity

The construct validity of a measure is another way of verifying the relationship between a particular measure and the concept it purportedly gauges. An instrument should operate systematically and in accord with an underlying construct. That is, persons who are inherently high on a particular skill or trait should perform accordingly on a test of this particular ability. Neuropsychological measures are proxies for neuropsychological constructs; these measures should be considered operational definitions of these very constructs. As mentioned earlier,

an operational definition is an explicit, quantifiable definition of the inclusion and exclusion criteria of a construct. For example, in developing a scale for aggression, researchers must decide whether to include verbal, physical, and/or passive aggressiveness into their definition. Once this is determined, the scale should assess only the relevant parameters of this construct.

As Bryant (2000) notes, face validity is somewhat related to construct validity, as it concerns the degree to which a measure appears to assess what it is intended. The immediate and delayed recall of information, for example, can be quite readily viewed as measures of learning and memory.

On the other hand, alternating between encircled numbers and letters on the Trail Making Test may not be such an obvious measure of executive cognition. Empirically, though, the Trail Making Test has been substantiated as a quite robust test of executive cognition. Similarly, self-ratings on the personality assessment inventory (PAI) (Morey, 1991) may be unrevealing to the naked eye. However, several items on the PAI tap peculiar and perhaps unlikely experiences that may suggest a psychogenic rather than an organic problem. However, the intent of these items is not (and should not be) apparent, and patients are generally unaware that clinically relevant (e.g., malingering versus psychosis) information is trying to be gleaned.

There are two broad criteria for assessing the construct validity of an instrument: convergent and divergent validity. The purpose of using such measures, often in combination, is to establish the specificity of a construct. Separate scales that tap into the same construct should be equally influenced. On the other hand, instruments that have no conceptual relation with a construct should not bear any statistical association.

Convergent Validity

Convergent validity examines whether instruments assessing the same construct "converge," or are in agreement. This merely means that if two measures are assessing the same underlying concept, they should behave similarly, and this evidently should manifest as a high correlation between the two instruments. For example, if performance on a test of activities of daily living (e.g., texas functional living scale; TFLS) (Cullum, Weiner, & Saine, 2009) is indeed an ecologically valid measure (i.e., samples a person's ability (or lack thereof) to perform similar activities out of the testing room), it should correlate highly with other valid measures: informant ratings of the person's activities of daily living; the number of errors they made cooking a meal at home, etc.

In developing a test of auditory/verbal learning and memory, high correlations should be demonstrated with other ("gold standards") tests that measure identical functions (e.g., Hopkins Verbal Learning Test—Revised) (Brandt & Benedict, 2001). High correlations between the two tests signify that these measures are presumably tapping into a unique dimension of cognition.

Divergent Validity

Divergent validity assesses whether measures of dissimilar constructs "diverge," in that they show no obvious relationship. This ensures both the integrity and the specificity of a construct.

If a new scale or depression was significantly related to another test of depression in a bivariate correlation, but this scale was completely distinct (uncorrelated) with a scale assessing anxiety, this would establish the test's divergent validity (as well as its convergent validity, since it is also correlated with another scale of depression).

Factor-analytic methods are common techniques for examining a measure's convergent and divergent validity simultaneously. Factor analysis, as discussed in this book, examines the covariation among a large number of items. Items that are highly correlated are subsumed under a particular factor, as they are assumed to be assessing the same latent variable or hypothetical construct. For example, to ensure the construct validity of a new instrument measuring a distinct component of depression, say social withdrawal, all individual items from this scale would be entered into the analysis. To corroborate social withdrawal's construct validity (i.e., construct validation), all items that relate to social withdrawal (e.g., preferring to stay at home rather than venture out, etc.) should load onto one factor. This establishes the convergent validity of the measure. On the other hand, if other items that tap into anhedonia and physiological symptoms of depression loaded separately, representing a distinct dimension of affective symptoms, this would provide evidence for social withdrawal's divergent validity. Together, the construct's convergent and divergent validity coalesce to establish construct validity.

There are other ways of determining a measure's divergent validity, such as simple bivariate correlations between measures. However, as convergent and divergent validity are often assessed concurrently, multivariate procedures are often preferred (e.g., Campbell & Fisk, 1959).

Clinical Validation

Bryant (2000) also discussed clinical validation in which researchers evaluate the accuracy with which scores on a given instrument can classify groups that are already known to differ on a criterion measure. For example, logistic regression analysis could be implemented to determine whether a screen for cognitive impairment differentiates persons with mild cognitive impairment and those healthy elderly persons (patients and healthy elderly persons represent the dichotomous criterion variable in the logistic regression analysis). This is clearly a subset of criterion validity, but it earns its moniker from the clinical context in which the data analysis arises.

Incremental Validity

Incremental validity holds a special place within clinical neuropsychology. The main goal of incremental validity is to determine whether the addition of potentially clinically relevant variables (e.g., test performance) contributes to a particular criterion measure over and above traditional tests. Statistically, this method involves examining the change in the proportion of variance explained in the dependent variable with the inclusion of an additional predictor. This is generally couched within a hierarchical regression analysis with the variable of primary importance placed in the last block (see chapter on regression).

It is important in neuropsychology to demonstrate that measures are able to improve classification or prediction above and beyond typical tests or measures used for diagnosis. This is exactly what we are looking at with incremental validity—how much more variance does a test, index, instrument explain *above and beyond* what is explained by other tests. More specifically, does a new test or measure improve an outcome measure beyond traditional tests? Does adding this measure help explain an outcome variable more so than if it had not been included?

As mentioned, hierarchical regression analysis is a commonly used method for examining the incremental validity of a test or measure. This holds true for either logistic regression (when you are classifying a binary outcome, such as normal controls and patients) or linear regression (when the outcome variable is continuous).

Busch, Frazier, Haggerty, and Kubu (2005) explored performance on the Boston Naming Test among 217 patients with intractable temporal lobe epilepsy (all right handed; 108 with left temporal lobe epilepsy and 109 with right temporal lobe epilepsy) and its ability to predict the ultimate side of surgery above and beyond the ability of indices of intellectual function (Verbal Comprehension Index and Perceptual Organization Index from the WAIS-III). In the hierarchical logistic regression analysis, scores from the WAIS-III along with a measure of delayed memory were entered into block/step 1. The raw score from the Boston Naming test was entered in block two. The dependent variable was the side of surgery (i.e., left or right).

To determine the significance of this model, the authors examined the last step (Step 2) to determine whether the raw score from the Boston Naming Test was significant. These authors examined the change in R^2 (proportion of variance) from models one (with just the scores from intelligence and memory) to model two (with the Boston Naming Test) to determine if this was a significant change in the prediction of side of surgery. The final model was significant, which supports the incremental validity of using the Boston Naming Test in classifying right versus left hemisphere surgery (among these particular tests).

Ecological Validity

Ecological validity has to do with how well neuropsychological tests are both conceptually and empirically related with activities of daily living (e.g., paying bills, driving, managing finances and medicines, misplacing one's keys, etc.).

Chaytor and Schmitter-Edgecombe (2003) recount the advent of technological advances (e.g., brain imaging) that have replaced certain goals of neuropsychological assessment. Brain imaging has supplanted neuropsychology from a practice of corroborating brain pathology (in most cases) to more broad applications that typically include the functioning of a person in whom pathology has been documented. Neuropsychologists are generally not counseled to opine on the mere presence of an organic etiology. This shift in roles for neuropsychology has required the profession to accommodate to an ever-changing environment. Simply put, brain imaging (whether structural or functional) will reveal nothing about a person's cognitive, behavioral, emotional, or functional capacities.

Chaytor and Schmitter-Edgecombe (2003) discuss two conceptual principles for establishing ecological validity: verisimilitude and veridicality. Verisimilitude concerns the correspondence and equivalency of tests and real-world demands. For example, tests are developed to simulate every day activities, such as writing a check, or using a telephone. Veridicality involves the degree to which a test shows an empirical relation to measures of everyday functioning (Franzen & Wilhelm, 1996). For example, a performance-based measure of daily activities—the texas functional living scale (TFLS) (Cullum et al., 2009)—should demonstrate a strong association with other aspects of real-world abilities, such as employment, or informant reports of the patient's activities of daily living.

Neuropsychological assessment remains weak at capturing all aspects that contribute to an individual's ability to function in the real world. There is a clear dissociation between a patient's performance on testing, and what he or she can do in his daily life. For one, unlike the real world, testing entails a distraction free, structured setting. There are also a host of noncognitive or nonintellectual factors that contribute to daily functioning. For example, personality and emotional difficulties, physical and sensory limitations, availability to certain resources (financial and environmental), living situations (e.g., house, apartment, dependence on car or public transportation, etc.) and many other perhaps ineffable factors affect, either for the better or worse, daily functioning (Chaytor & Schmitter-Edgecombe, 2003).

Summary

This chapter discussed validity as it pertains to neuropsychology. Five principal aspects of validity were delineated: Content, criterion, construct, incremental, and ecological. Such methods vary in their conceptual and practical goals, but show considerable overlap in terms of the statistical models that attempt to establish their veracity.

Whereas content validity ensures that the items used on a test are theoretically culled from a population of relevant items to ensure exhaustiveness in its assessment, methods for criterion validity (postdictive, concurrent, and predictive) vary predominantly in the temporal order in which the variables are assessed. The goal of criterion validity is to ensure that tests we use as proxies for certain behaviors are associated, and presumably, can predict a future outcome. Construct validity

(convergent and divergent) ensures that the constructs we use in neuropsychology are indeed evaluating what we say they are. Incremental validity helps improve prediction of additional tests and constructs by demonstrating their unique contribution to classifying particular outcomes over and above standard methods. And, ecological validity attempts to bridge the gap between what is exhibited on neuropsychological testing and what is experienced in the real world. That is, how well does performance on neuropsychological testing reflect one's daily functioning? Does someone who has trouble with various activities of daily living show commensurate difficulty when examined in an office by a neuropsychologist?

Overall, validity is a process of establishing a test's clinical utility in describing, explaining, and predicting phenomena. Validity is not a function of a test per se, but is inextricably entwined with the context in which a test is applied.

http://www.springer.com/978-1-4614-3416-0

Statistical Methods in Neuropsychology
Common Procedures Made Comprehensible
Maroof, D.A.
2012, X, 110 p., Hardcover
ISBN: 978-1-4614-3416-0