Encyclopedia of Research Design

Criterion Validity

Contributors: Matthew J. Borneman
Editors: Neil J. Salkind
Book Title: Encyclopedia of Research Design
Chapter Title: "Criterion Validity"
Pub. Date: 2010
Access Date: October 14, 2013
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412961271
Online ISBN: 9781412961288
DOI: http://dx.doi.org/10.4135/9781412961288.n88
Print pages: 292-297

This PDF has been generated from SAGE Research Methods. Please note that the pagination of the online version will vary from the pagination of the print book.

http://dx.doi.org/10.4135/9781412961288.n88

Also known as *criterion-related validity*, or sometimes *predictive* or *concurrent validity, criterion validity* is the general term to describe how well scores on one measure (i.e., a *predictor*) predict scores on another measure of interest (i.e., the *criterion*). In other words, a particular criterion or outcome measure is of interest to the researcher; examples could include (but are not limited to) ratings of job performance, grade point average (GPA) in school, a voting outcome, or a medical diagnosis. *Criterion validity*, then, refers to the strength of the relationship between measures intended to predict the ultimate criterion of interest and the criterion measure itself. In academic settings, for example, the criterion of interest may be GPA, and the predictor being studied is the score on a standardized math test. Criterion validity, in this context, would be the strength of the relationship (e.g., the correlation coefficient) between the scores on the standardized math test and GPA.

Some care regarding the use of the term *criterion validity* needs to be employed. Typically, the term is applied to predictors, rather than criteria; researchers often refer to the “criterion validity” of a specific predictor. However, this is not meant to imply that there is only one “criterion validity” estimate for each predictor. Rather, each predictor can have different “criterion validity” estimates for many different criteria. Extending the above example, the standardized math test may have one criterion validity estimate for overall GPA, a higher criterion validity estimate for science ability, and a lower criterion validity estimate for artistic appreciation; all three are valid criteria of interest. Additionally, each of these estimates may be moderated by (i.e., have different criterion validity estimates for) situational, sample, or research design characteristics. In this entry the criterion, the research designs that assess criterion validity, effect sizes, and concerns that may arise in applied selection are discussed.

Nature of the Criterion

Again, the term *criterion validity* typically refers to a *specific* predictor measure, often with the criterion measure assumed. Unfortunately, this introduces substantial confusion into the procedure of criterion validation. Certainly, a single predictor measure can predict an extremely wide range of criteria, as Christopher Brand has shown with general intelligence, for example. Using the same example, the criterion validity estimates for general intelligence vary quite a bit; general intelligence predicts some criteria better than others. This fact further illustrates that there is no single criterion validity estimate for a single predictor. Additionally, the relationship between one predictor measure and one criterion variable can vary depending on other variables (i.e., moderator variables), such as situational characteristics, attributes of the sample, and particularities of the research design. Issues here are highly related to the criterion problem in predictive validation studies.

Research Design

There are four broad research designs to assess the criterion validity for a specific predictor: predictive validation, quasi-predictive validation, concurrent validation, and postdictive validation. Each of these is discussed in turn.

Predictive Validation

When examining the criterion validity of a specific predictor, the researcher is often interested in selecting persons based on their scores on a predictor (or set of predictor measures) that will predict how well the people will perform on the criterion measure. In a true predictive validation design, predictor measure or measures are administered to a set of applicants, and the researchers select applicants completely randomly (i.e., without regard to their scores on the predictor measure or measures.) The correlation between the predictor measure(s) and the criterion of interest is the *index of criterion validity.* This design has the advantage of being free **[p. 292 ↓]** from the effects of range restriction; however, it is an expensive design, and unfeasible in many situations, as stakeholders are often unwilling to forgo selecting on potentially useful predictor variables.

Quasi-Predictive Validation

Like a true predictive validation design, in a quasi-predictive design, the researcher is interested in administering a predictor (or set of predictors) to the applicants in order to predict their scores on a criterion variable of interest. Unlike a true predictive design, in a quasi-predictive validation design, the researcher will select applicants based on their scores on the predictor(s). As before, the correlation between the predictor(s) and the criterion of interest is the index of criterion validity. However, in a quasi-predictive design, the correlation between the predictor and criterion will likely be smaller because of range restriction due to selection on the predictor variables. Certainly, if the researcher has a choice between a predictive and quasi-predictive design, the predictive design would be preferred because it provides a more accurate estimate of the criterion validity of the predictor(s); however, quasi-predictive designs are far more common. Although quasi-predictive designs typically suffer from range restriction problems, they have the advantage of allowing the predictors to be used for selection purposes while researchers obtain criterion validity estimates.

Concurrent Validation

In a concurrent validation design, the predictor(s) of interest to the researcher are not administered to a set of applicants; rather, they are administered only to the incumbents, or people who have already been selected. The correlation between the scores on the predictors and the criterion measures for the incumbents serves as the criterion validity estimate for that predictor or set of predictors. This design has several advantages, including cost savings due to administering the predictors to fewer people and reduced time to collection of the criterion data. However, there are also some disadvantages, including the fact that criterion validity estimates are likely to be smaller as a result of range restriction (except in the rare situation when the manner in which the incumbents were selected is completely unrelated to scores on the predictor or predictors).

Another potential concern regarding concurrent validation designs is the motivation of test takers. This is a major concern for noncognitive assessments, such as personality tests, survey data, and background information. Collecting data on these types of assessments in a concurrent validation design provides an estimate of the maximum criterion validity for a given assessment. This is because incumbents, who are not motivated to alter their scores in order to be selected, are assumed to be answering honestly. However, there is some concern for intentional distortion in motivated testing sessions (i.e., when applying for a job or admittance to school), which can affect criterion validity estimates. As such, one must take care when interpreting criterion validity estimates in this type of design. If estimates under operational selection settings are of interest (i.e., when there is some motivation for distortion), then criterion validity estimates from a predictive or quasi-predictive design are of interest; however, if estimates of maximal criterion validity for the predictor(s) are of interest, then a concurrent design is appropriate.

Postdictive Validation

Postdictive validation is an infrequently used design to assess criterion validity. At its basics, postdictive validation assesses the criterion variable first and then subsequently assesses the predictor variable(s). Typically, this validation design is not employed because the predictor variable(s), by definition, come temporally before the criterion variable is assessed. However, a postdictive validation design can be especially useful, if not the only alternative, when the criterion variable is rare or unethical to obtain. Such examples might include criminal activity, abuse, or medical outcomes. In rare criterion instances, it is nearly impossible to know when the outcome will occur; as such, the predictors are collected after the fact to help predict who is at risk for the particular criterion variable. In other instances when it is extremely unethical to collect data on the criterion of interest (e.g., abuse), predictor variables are collected after the fact in order to determine who might be at risk for those criterion variables. Regardless of the **[p. 293 ↓]** reason for the postdictive design, people who met or were assessed on the criterion variable are matched with other people who were not, typically on demographic and/or other variables. The relationship between the predictor measures and the criterion variable assessed for the two groups serves as the estimate of criterion validity.

Effect Sizes

Any discussion of criterion validity necessarily involves a discussion of effect sizes; the results of a statistical significance test are inappropriate to establish criterion validity. The question of interest in criterion validity is, To what degree are the predictor and criterion related? or How well does the measure predict scores on the criterion variable? instead of, Are the predictor and criterion related? Effect sizes address the former questions, while significance testing addresses the latter. As such, effect sizes are necessary to quantify how well the predictor and criterion are related and to provide a way to compare the criterion validity of several different predictors.

The specific effect size to be used is dependent on the research context and types of data being collected. These can include (but are not limited to) odds ratios, correlations, and standardized mean differences. For the purposes of explanation, it is assumed that there is a continuous predictor and a continuous criterion variable, making the correlation coefficient the appropriate measure of effect size. In this case, the correlation between a given predictor and a specific criterion serves as the estimate of criterion validity. Working in the effect size metric has the added benefit of permitting comparisons of criterion validity estimates for several predictors. Assuming that two predictors were collected under similar research designs and conditions and are correlated with the same criterion variable, then the predictor with the higher correlation with the criterion can be said to have greater criterion validity than the other predictor (for that particular criterion and research context). If a criterion variable measures different behaviors or was collected under different research contexts (e.g., a testing situation prone to motivated distortion vs. one without such motivation), then criterion validity estimates are not directly comparable.

Statistical Artifacts

Unfortunately, several statistical artifacts can have dramatic effects on criterion validity estimates, with two of the most common being measurement error and range restriction. Both of these (in most applications) serve to lower the observed relationships from their true values. These effects are increasingly important when one is comparing the criterion validity of multiple predictors.

Range Restriction

Range restriction occurs when there is some mechanism that makes it more likely for people with higher scores on a variable to be selected than people with lower scores. This is common in academic or employee selection as the scores on the administered predictors (or variables related to those predictors) form the basis of who is admitted or hired. Range restriction is common in quasi-predictive designs (because predictor scores are used to select or admit people) and concurrent designs (because people are selected in a way that is related to the predictor variables of interest in the study). For example, suppose people are hired into an organization on the basis of their interview scores. The researcher administers another potential predictor of the focal criterion in a concurrent validation design. If the scores on this new predictor are correlated with scores on the interview, then range restriction will occur. True predictive validation designs are free from range restriction because either no selection occurs or selection occurs in a way uncorrelated with the predictors. In postdictive validation designs, any potential range restriction is typically controlled for in the matching scenario.

Range restriction becomes particularly problematic when the researcher is interested in comparing criterion validity estimates. This is because observed criterion validity estimates for different predictors can be differentially decreased because of range restriction. Suppose that two predictors that truly have equal criterion validity were administered to a set of applicants for a position. Because of the nature of the way they were selected, suppose that for Predictor A, 90% of the variability in predictor scores remained after people were selected, but only 50% of the variability remained for Predictor B after selection. Because **[p. 294 ↓]** of the effects of range restriction, Predictor A would have a higher criterion validity estimate than Predictor B would, even though each had the same true criterion validity. In these cases, one should apply range restriction corrections before comparing validity coefficients. Fortunately, there are multiple formulas available to correct criterion validity estimates for range restriction, depending on the precise mechanism of range restriction.

Measurement Error

Unlike range restriction, the attenuation of criterion validity estimates due to unreliability occurs in all settings. Because no measure is perfectly reliable, random measurement error will serve to attenuate statistical relationships among variables. The well-known correction for attenuation serves as a way to correct observed criterion validity estimates for attenuation due to unreliability. However, some care must be taken in applications of the correction for attenuation.

In most applications, researchers are not interested in predicting scores on the criterion *measure;* rather, they are interested in predicting standing on the criterion *construct.* For example, the researcher would not be interested in predicting the supervisory ratings of a particular employee's teamwork skills, but the researcher would be interested in predicting the true nature of the teamwork skills. As such, correcting for attenuation due to measurement error in the criterion provides a way to estimate the relationship between predictor scores and the true criterion construct of interest. These corrections are extremely important when criterion reliability estimates are different in the validation for multiple predictors. If multiple predictors are correlated with the same criterion variable in the same sample, then each of these criterion validity estimates is attenuated to the same degree. However, if different samples are used, and the criterion reliability estimates are unequal in the samples, then the correction for attenuation in the criterion should be employed before making comparisons among predictors.

Corrections for attenuation in the predictor variable are appropriate only under some conditions. If the researcher is interested in a theoretical relationship between a predictor *construct* and a criterion construct, then correcting for attenuation in the predictor is warranted. However, if there is an applied interest in estimating the relationship between a particular predictor *measure* and a criterion of interest, then correcting for attenuation in the predictor is inappropriate. Although it is true that differences in predictor reliabilities can produce artifactual differences in criterion validity estimates, these differences have substantive implications in applied settings. In these instances, every effort should be made to ensure that the predictors are as reliable as possible for selection purposes.

Concerns in Applied Selection

In applied purposes, researchers are often interested not only in the criterion validity of a specific predictor (which is indexed by the appropriate effect size) but also in predicting scores on the criterion variable of interest. For a single predictor, this is done with the equation

$$y_i = b_0 + b_1 x_i, \qquad (1)$$

where y_i is the score on the criterion variable for person i, x_i is the score on the predictor variable for person i, b_0 is the intercept for the regression model, and b_1 is the slope for predictor x. Equation 1 allows the researcher to predict scores on the criterion variable from scores on the predictor variable. This can be especially useful when a researcher wants the performance of selected employees to meet a minimum threshold.

When multiple predictors are employed, the effect size of interest is not any single bivariate correlation but the multiple correlation between a set of predictors and a single criterion of interest (which might be indexed with the multiple R or R^2 from a regression model). In these instances, the prediction equation analogous to Equation 1 is

$$y_i = b_0 + b_{1x1i} + b_{2x2i} + \cdots + b_p x_{pi}, \qquad (2)$$

where x_{1i}, x_{2i}, … x_{pi} are the scores on the predictor variables 1, 2, … *p* for person *i*, b_1, b_2, … b_p are the slopes for predictors x_1, x_2, … x_p, and other terms are as defined earlier. Equation 2 allows the researcher to predict scores on **[p. 295 ↓]** a criterion variable given scores on a set of *p* predictor variables.

Predictive Bias

A unique situation arises in applied selection situations because of federal guidelines requiring criterion validity evidence for predictors that show adverse impact between *protected groups.* Protected groups include (but are not limited to) ethnicity, gender, and age. Adverse impact arises when applicants from one protected group (e.g., males) are selected at a higher rate than members from another protected group (e.g., females). Oftentimes, adverse impact arises because of substantial group differences on the predictor on which applicants are being selected. In these instances, the focal predictor must be shown to exhibit criterion validity across all people being selected. However, it is also useful to examine predictive bias.

For the sake of simplicity, predictive bias will be explicated here only in the case of a single predictor, though the concepts can certainly be extended to the case of multiple predictors. In order to examine the predictive bias of a criterion validity estimate for a specific predictor, it is assumed that the variable on which bias is assessed is categorical; examples would include gender or ethnicity. The appropriate equation would be

$$y_i = b_0 + b_{1x1i} + b_{2x2i} + b_3(x_{1i} * x_{2i}), \qquad (3)$$

where x_{1i} and x_{2i} are the scores on the continuous predictor variable and the categorical demographic variable, respectively, for person *i*, b_1 is the regression coefficient for the continuous predictor, b_2 is the regression coefficient for the categorical predictor, b_3 is the regression coefficient for the interaction term, and other terms are defined as earlier. Equation 3 has substantial implications for bias in criterion validity estimates.

Assuming the reference group for the categorical variable (e.g., males) is coded as 0 and the focal group (e.g., females) is coded as 1, the b_0 coefficient gives the intercept for the reference group, and the b_1 coefficient gives the regression slope for the reference group. These two coefficients form the baseline of criterion validity evidence for a given predictor. The b_2 coefficient and the b_3 coefficient give estimates of how the intercept and slope estimates, respectively, change for the focal group.

The b_2 and b_3 coefficients have strong implications for bias in criterion validity estimates. If the b_3 coefficient is large and positive (negative), then the slope differences (and criterion validity estimates) are substantially larger (smaller) for the focal group. However, if the b_3 coefficient is near zero, then the criterion validity estimates are approximately equal for the focal and reference groups. The magnitude of the b_2 coefficient determines (along with the magnitude of the b_3

coefficient) whether the criterion scores are over- or underestimated for the focal or references groups depending on their scores on the predictor variable. It is generally accepted that for predictor variables with similar levels of criterion validity, those exhibiting less predictive bias should be preferred over those exhibiting more predictive bias. However, there is some room for tradeoffs between criterion validity and predictive bias.

Matthew J. Borneman

http://dx.doi.org/10.4135/9781412961288.n88
See also

Further Readings

Binning, J. F., & and Barrett, G. V. Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied Psychology, (1989). vol. 74, pp. 478–494.

Brand, C. (1987). The importance of general intelligence. In S. Modgil, ed. & C. Modgil (Eds.), Arthur Jensen: Consensus and controversy (pp. pp. 251–265). Philadelphia: Falmer Press.

Cleary, T. A. Test bias: Prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement, (1968). vol. 5, pp. 115–124.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). Measurement theory for the behavioral sciences. San Francisco: W. H. Freeman.

Kuncel, N. R., & and Hezlett, S. A. Standardized tests predict graduate students' success. Science, (2007). vol. 315, pp. 1080–1081.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Sackett, P. R.,, Schmitt, N.,, Ellingson, J. E., &, and Kabin, M. B. High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. American Psychologist, (2001). vol. 56, pp. 302–318.

Schmidt, F. L., & and Hunter, J. E. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, (1998). vol. 124, pp. 262–274.