Journal of Personality and Social Psychology
2003, Vol. 84, No. 3, 608–618

Copyright 2003 by the American Psychological Association, Inc.
0022-3514/03/$12.00 DOI: 10.1037/0022-3514.84.3.608

Quantifying Construct Validity: Two Simple Measures

Drew Westen
Emory University

Robert Rosenthal
University of California, Riverside and Harvard University

Construct validity is one of the most central concepts in psychology. Researchers generally establish the construct validity of a measure by correlating it with a number of other measures and arguing from the pattern of correlations that the measure is associated with these variables in theoretically predictable ways. This article presents 2 simple metrics for quantifying construct validity that provide effect size estimates indicating the extent to which the observed pattern of correlations in a convergent-discriminant validity matrix matches the theoretically predicted pattern of correlations. Both measures, based on contrast analysis, provide simple estimates of validity that can be compared across studies, constructs, and measures meta-analytically, and can be implemented without the use of complex statistical procedures that may limit their accessibility.

> The best construct is the one around which we can build the greatest number of inferences, in the most direct fashion. (Cronbach & Meehl, 1955, p. 288)

Construct validity is one of the most important concepts in all of psychology. It is at the heart of any study in which researchers use a measure as an index of a variable that is not itself directly observable (e.g., intelligence, aggression, working memory). If a psychological test (or, more broadly, a psychological procedure, including an experimental manipulation) lacks construct validity, results obtained using this test or procedure will be difficult to interpret. Not surprisingly, the "construct" of construct validity has been the focus of theoretical and empirical attention for over half a century, especially in personality, clinical, educational, and organizational psychology, where measures of individual differences of hypothesized constructs are the bread and butter of research (Anastasi & Urbina, 1997; Cronbach & Meehl, 1955; Nunnally & Bernstein, 1994).

Yet, despite the importance of this concept, no simple metric can be used to quantify the extent to which a measure can be described as construct valid. Researchers typically establish construct validity by presenting correlations between a measure of a construct and a number of other measures that should, theoretically, be associated with it (convergent validity) or vary independently of it (discriminant validity). Thus, a researcher interested in "rumination" as a personality trait might present correlations indicating that her new measure correlates $r = .45$ with the Neuroticism factor of the Revised NEO Personality Inventory (NEO-PI-R; McCrae & Costa, 1987) and $r = .60$ with Trait Anxiety but correlates only modestly ($r = .10$) with the Openness factor of the NEO-PI-R and negatively with a measure of Active Coping ($r = -.30$). Appealing to the seemingly sensible pattern of correlations (e.g., ruminators should be anxious, and their rumination is likely to interfere with active coping efforts but should be orthogonal to openness), the researcher concludes that she has accrued evidence, at least in a preliminary way, for the construct validity of her new trait and measure, and she goes on to conduct a series of programmatic studies thereafter, based on the complex pattern of convergent and discriminant validity coefficients (*r*s) that help define the construct validity of the measure.

The aim of construct validation is to embed a purported measure of a construct in a nomological network, that is, to establish its relation to other variables with which it should, theoretically, be associated positively, negatively, or practically not at all (Cronbach & Meehl, 1955). A procedure designed to help quantify construct validity should provide a summary index not only of whether the measure correlates positively, negatively, or not at all with a series of other measures, but the relative magnitude of those correlations. In other words, it should be an index of the extent to which the researcher has accurately predicted the pattern of findings in the convergent-discriminant validity array. Such a metric should also provide a test of the statistical significance of the match between observed and expected correlations, and provide confidence intervals for that match, taking into account the likelihood that some of the validating variables may not be independent of one another.

In this article we present two effect size estimates (correlation coefficients) for quantifying construct validity, designed to summarize the pattern of findings represented in a convergent-discriminant validity array for a given measure. These metrics provide simple estimates of validity that can be compared across studies, constructs, and measures. Both metrics provide a quantified index of the degree of convergence between the observed pattern of correlations and the theoretically predicted pattern of correlations—that is, of the degree of agreement of the data with the theory underlying the construct and the measure.

Drew Westen, Department of Psychology and Department of Psychiatry and Behavioral Sciences, Emory University; Robert Rosenthal, Department of Psychology, University of California, Riverside and Department of Psychology, Harvard University.

Preparation of this article was supported in part by National Institute of Mental Health Grants MH59685 and MH60892 to Drew Westen.

Software for computing the construct validity coefficients described in this article will be available at www.psychsystems.net/lab

Correspondence concerning this article should be addressed to Drew Westen, Department of Psychology and Department of Psychiatry and Behavioral Sciences, Emory University, 532 North Kilgo Circle, Atlanta, Georgia 30322. E-mail: dwesten@emory.edu

Construct Validation

Over the past several decades psychologists have gradually refined a set of methods for assessing the validity of a measure. In broadest strokes, psychologists have distinguished a number of kinds of statements about the validity of a measure, including (a) content validity, which refers to the extent to which the measure adequately samples the content of the domain that constitutes the construct (e.g., different behavioral expressions of rumination that should be included in a measure of rumination as a personality trait); (b) criterion validity, which refers to the extent to which a measure is empirically associated with relevant criterion variables, which may either be assessed at the same time (concurrent validity), in the future (predictive validity), or in the past (postdictive validity); and (c) construct validity, an overarching term now seen by most to encompass all forms of validity, which refers to the extent to which a measure adequately assesses the construct it purports to assess (Nunnally & Bernstein, 1994).

Two points are important to note here about construct validity. First, although researchers often describe their instruments as "validated," construct validity is an estimate of the extent to which variance in the measure reflects variance in the underlying construct. Virtually all measures (except those using relatively infallible indicators, such as measures of biological sex) include error components that reflect not only random factors but method variance (variance attributable to the method being used, such as self-report vs. interviews) and irrelevant but nonrandom variables that have been inadvertently included in the measure. Thus, a researcher studying rumination would want to show that the measure correlates with trait anxiety, but would also want to demonstrate that something is left when holding anxiety constant other than random error and method variance—that is, something unique to rumination over and above anxiety.

Second, construct validation is always theory dependent (Cronbach & Meehl, 1955). A statement about the validity of an instrument is a statement about the extent to which its observed associations with measures of other variables match theoretical predictions about how it should be associated with those variables. If the theory is wrong, the pattern of correlations will appear to invalidate the measure. Construct validation is a bootstrapping operation: Initial (often vague and intuitive) theories about a construct lead to creation of a measure designed to have content validity vis-à-vis the construct as understood at that point in time (Cronbach & Meehl, 1955). Subsequently, researchers assess the relation between the measure and relevant criterion variables and determine the extent to which (a) the measure needs to be refined, (b) the construct needs to be refined, or (c) more typically, both. Thus, construct validation is not only continuous (a matter of degree, not a categorical distinction between valid and invalid) but continual (a perpetual, self-refining process).

Contrast Analysis and Construct Validity

In their classic article on construct validation, Cronbach and Meehl (1955) considered the possibility of developing an overall coefficient for indexing construct validity but noted the difficulty of providing anything more than a broad indication of the upper and lower bounds of validity. However, developments since that time, particularly in the concept of the multitrait–multimethod matrix (Campbell & Fiske, 1959; Shrout & Fiske, 1995), have led to continued efforts to derive more quantitative, less impressionistic ways to index the extent to which a measure is doing its job. Thus, a number of researchers have developed techniques to try to separate out true variance on a measure of a trait from method variance, often based on the principle that method effects and trait effects (and their interactions) should be distinguishable using analysis of variance (ANOVA), confirmatory factor analysis (because trait and method variance should load on different factors), structural equation modeling (SEM), and related statistical procedures (Cudeck, 1988; Hammond, Hamm, & Grassia, 1986; Kenny, 1995; Reichardt & Coleman, 1995; Wothke, 1995).

The procedure we describe here is in many respects similar, but is simple, readily applied, and designed to address the most common case in which a researcher wants to validate a single measure by correlating it with multiple other measures. In putting forth this method, we are not suggesting that researchers should avoid using other techniques, for example, SEM, which may be well suited to modeling complex relationships and can produce quite elegant portraits of the kinds of patterns of covariation that constitute multitrait–multimethod matrices. However, no approach has yet gained widespread acceptance or been widely used to index construct validity, and we believe the present approach has several advantages.

First, it accords with a primary tenet of the American Psychological Association's Task Force on Statistical Inference (Wilkinson & Task Force on Statistical Inference, 1999), which is to use minimally sufficient statistics and to avoid more complex analyses when simpler ones will do. As described below, the procedures outlined here require the researcher to specify in advance a predicted pattern of correlations between a measure of a construct and criterion variables, nothing more. The researcher need not estimate or make assumptions about any parameters other than those under investigation.

Second and related, SEM requires assumptions that are unnecessary using the procedures outlined here, which rely instead on simple product-moment correlations. Indeed, one of the reasons we believe approaches based on SEM have not been widely used for purposes of construct validation by personality and clinical researchers despite their elegance is their dependence on particular characteristics of the data (and typically large sample sizes) that are not always present.

Third, the approach we propose, based on contrast analysis, asks a highly specific, focused question with one degree of freedom. The question it addresses is whether the researcher has accurately predicted the magnitude of correlations between a single predictor variable and multiple criterion variables. Rosenthal, Rosnow, and Rubin (2000) have outlined the advantages of focused questions of this sort, but the major advantage worth emphasizing here is that these procedures based on one degree of freedom provide a single answer to a single question (in this case, does this measure predict an array of other measures in a way predicted by theory?). SEM, in contrast, uses significance tests of $df > 1$ and is specifically designed to answer multiple questions simultaneously.

Fourth, the method proposed here yields simple, readily understood indices (product-moment correlations). Product-moment correlations are among the most common and interpretable effect size estimates, and any effect size estimate can be converted to Pearson's r (Rosenthal & DiMatteo, 2001). Researchers can readily interpret the meaning of an r by converting it to a binomial effect size display, which displays in percentages the practical benefits of particular correlations as they depart from zero more and more (Rosenthal & Rubin, 1982); or by consulting a table of correlations between variables whose relationship is well established, such as sex differences in aggression or the relation between average temperature and distance from sea level (Meyer et al., 2001). What makes a summary index in the form of a simple *r* even more useful is that the index and the variables that comprise it are in precisely the same form. When researchers describe the construct validity of a measure, they index its relation to other variables using *r*. A summary statistic that could combine these individual correlations into a single *r* would provide a meaningful and readily interpretable index.

Finally, the *r* produced in one study is directly comparable with the *r* produced in another, so that data can be easily aggregated across studies meta-analytically. Although SEM produces goodness-of-fit indices, these can be difficult to interpret for those who do not deal regularly with the procedure, particularly in the absence of clear benchmarks familiar to most users of psychological tests. Goodness-of-fit indices are also difficult to aggregate across studies quantitatively, based as they are on often widely varying degrees of freedom, and they summarize the fit of an entire set of equations taken simultaneously (including parameter estimates) rather than providing a specific fit index between a single measure and a set of other measures with which it is expected to correlate in particular ways.

The procedure we are proposing derives primarily from recent developments in contrast analysis (Meng, Rosenthal, & Rubin, 1992; Rosenthal et al., 2000), a set of techniques usually used in ANOVA to test specific hypotheses about the relative magnitude of a series of means. For example, using contrast analysis to examine group differences in rumination, a researcher might predict that the mean of Group A (anxious patients) will be higher than the means of Groups B and C (psychopaths and normals, respectively). This prediction would be tested by selecting contrast weights (lambda weights) that reflect the predicted ordering of the means—in this case, 2, −1, −1. Any values representing the predicted relative magnitudes of means (or of other statistics) can be selected as contrast weights, but for any contrast, these weights must sum to zero. Our rumination researcher might also predict a linear relation among the three groups, such that anxious patients should be highest on rumination, followed by normals, followed by psychopaths (who, according to many theories, should be particularly low on anxiety), and hence assign the contrast weights +1, 0, and −1. These contrasts are essentially a generalization from the common two-group comparison (*t* test), in which the contrast weights are +1 and −1. The correlation coefficient, *r*, provides a useful effect size estimate of the contrast in the two-group case (the correlation between presence or absence of the independent variable, coded 0/1, and the dependent variable), and can similarly be derived for a comparison of any number of groups.

Contrast analysis generally yields an effect size estimate (*r*), an associated significance test (e.g., *t* or *Z*), and a *p*-value. It provides an answer to a highly focused question (i.e., How well does the predicted pattern of means among the groups resemble the observed pattern?) rather than to an unfocused question (i.e., Do the means across the *k* groups differ from one another in some way?). Thus, unlike an omnibus *F* statistic, contrast analysis allows the investigator to test a particular hypothesis, specified in advance, about where and how group differences should occur. Doing so substantially increases statistical power.

Although researchers have most commonly applied this method to ANOVA in experimental designs, contrast analysis is equally applicable to correlational data. Just as researchers can construct contrasts to test the relative ordering of means, they can equally construct contrasts to assess the relative ordering of correlation coefficients, even when those correlation coefficients are correlated with one another (Meng et al., 1992; Rosenthal et al., 2000).

Two Construct Validity Coefficients: $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$

Two effect size correlations provide convenient and informative indices of construct validity, each in its own way. The first of these correlations, $r_{alerting\text{-}CV}$, is the simple correlation between (a) the pattern of correlations *predicted* between the measure being validated and the *k* variables correlated with that measure, and (b) the pattern of correlations actually *obtained*. It is called an “alerting” correlation because it is a rough, readily interpretable index that can alert the researcher to possible trends of interest (Rosenthal et al., 2000).

For example, suppose we were developing a new measure of social intelligence. We have administered our new measure to a sample of individuals to whom we have also administered five other measures. Our construct of social intelligence is such that we predict it will correlate with the five other measures as follows: (a) Verbal IQ, *r* predicted as .5; (b) Nonverbal decoding skill, *r* predicted as .4; (c) Extraversion, *r* predicted as .1; (d) Minnesota Multiphasic Interventory (MMPI-2) Psychopathic Deviate Scale, *r* predicted as −.1; and (e) Hostile attributional bias, *r* predicted as −.4. To compute $r_{alerting\text{-}CV}$, we simply correlate these predicted values (arranged as a column of data) with the obtained values (arranged as a second column of data). More accurate results are obtained when the correlations (*r*s) are first transformed into their Fisher Z_r equivalents in order to improve normality (Meng et al., 1992; Steiger, 1980).

Thus, suppose the obtained values (Z_r transformed) were *r*s of .30, .10, .20, .00, and −.20. The correlation between this column of data and our predicted values (.5, .4, .1, −.1, −.4) yields an $r_{alerting\text{-}CV}$ of .88. The magnitude of this correlation suggests that our predicted pattern of values provided a very accurate portrayal of the pattern or profile of correlations actually obtained.

The effect size correlation $r_{alerting\text{-}CV}$ becomes increasingly useful as we include more criterion variables in our convergent-discriminant validity matrix. If only two variables are to be correlated with our new measure, $r_{alerting\text{-}CV}$ can take on values of only +1.00 or −1.00. As more variables are added, $r_{alerting\text{-}CV}$ becomes more informative. To put it another way, $r_{alerting\text{-}CV}$ provides an unstable index when the number of criterion variables is small but becomes progressively more useful as the researcher makes bolder hypotheses about the relation between the target measure and a range of criterion variables—that is, as the nomological net gets wider. We typically do not compute *p* levels for $r_{alerting\text{-}CV}$, but it can be used to help in the computation of significance levels for our other effect size correlation, $r_{contrast\text{-}CV}$ (see Equation A5 in Appendix A).

Our second correlation, $r_{contrast\text{-}CV}$, shares with $r_{alerting\text{-}CV}$ the characteristic that it will be larger as the match between expected and obtained correlations is higher. In addition, however, $r_{contrast\text{-}CV}$ uses information about (a) the median intercorrelation among the variables to be correlated with the measure being validated, and (b) the absolute values of the correlations between the measure being validated and the variables with which it is being correlated. A felicitous feature of $r_{contrast\text{-}CV}$ is that the interpretation of this second metric is not limited in the same way as is $r_{alerting\text{-}CV}$ when there are only a few variables in the convergent-discriminant validity matrix. Computational details for $r_{contrast\text{-}CV}$ are provided in Appendix A.

Effect size correlations such as $r_{contrast\text{-}CV}$ are typically seen in a context of comparing a set of group means. Suppose we were examining the effects on performance of four levels of medication—for example, 100, 200, 300, and 400 mg—and found mean performance scores of the four groups to be 6, 8, 10, and 12, respectively, with a standard deviation of 5 and a sample size of 10 in each of the groups. Whereas $r_{alerting}$, the correlation between the contrast weights of -3, -1, $+1$, and $+3$ (for the four equally spaced levels of medication) and their corresponding means of 6, 8, 10, and 12 would be 1.00, the $r_{contrast}$ would be only .43. We can think of $r_{contrast}$, in this example, as the correlation between the contrast weights and the individual scores of the 40 patients. The greater the variability within conditions, the lower the $r_{contrast}$. This is not the case for $r_{alerting}$, which is unaffected by the variability within conditions because it is based only on the condition means (Rosenthal et al., 2000).

Just as effect size correlations such as $r_{alerting}$ and $r_{contrast}$ can be used to compare *means* with a theory-based set of predictions (lambda weights, or λs), they can also be used to compare *correlations* with a theory-based set of predictions (also expressed as λs). The use of $r_{contrast\text{-}CV}$ differs somewhat from the more common use of $r_{contrast}$ in ANOVA in that group means are typically independent of each other, whereas the correlations used in $r_{contrast\text{-}CV}$ are typically not independent of each other.

When researchers apply contrast analysis to correlations that are independent, as in meta-analytic work, very simple equations can be used (Rosenthal, 1991). When they compare correlations that are not independent, however, the median intercorrelation among the variables to be examined is required. As this median intercorrelation increases, for a given degree of agreement between predicted and obtained correlations, so usually does the $r_{contrast\text{-}CV}$. That is because it is harder to achieve any given level of agreement when the variables are more similar to each other (as indexed by a large median intercorrelation). The other ingredient required when the correlations being examined are not independent is the average squared correlation between the measure being validated and the criterion variables. As this average squared correlation decreases to near .00 or increases to near 1.00, $r_{contrast}$ usually tends to increase because it is harder to achieve any given level of agreement when there is minimal variation among the correlations for which differential predictions have been made.

An Example

To provide an example, we describe actual data recently collected on adolescent personality pathology. Respondents were 266 randomly selected psychiatrists and psychologists with an average of roughly 15 years experience who were asked to describe a randomly selected patient they were currently treating between the ages of 14 and 18 years for "enduring, maladaptive patterns of thought, feeling, motivation, and behavior—that is, personality" (Westen & Chang, 2000; Westen, Shedler, Durrett, Glass, & Martens, in press). The aim of the study was to develop an empirically grounded classification of adolescent personality pathology that does not assume current Axis II categories and criteria developed from observation of adults. Clinicians provided a description of the patient's personality using a 200-item Q-sort. The investigators then applied an empirical clustering technique, Q-analysis, to these data to try to discern whether patients naturally fall into groupings (categories or dimensions) based on their personality profiles. These empirically derived personality styles then provided prototypes against which each patient's profile could be compared. Doing so yielded dimensional scores assessing the degree of match between the patient's personality and each of the empirically derived prototypes (much like an MMPI-2 profile reflects the match between a patient's responses and the responses of criterion groups). Thus, converting to T-scores, a patient might receive a score of 70 on antisocial–psychopathic (one of the diagnoses that emerged empirically), 45 on narcissistic (another diagnosis identified by Q-analysis), 32 on histrionic, and so forth.

The next step was to try to validate these new dimensions—to begin to locate them within a nomological network. For the present purposes, we randomly selected one of these dimensions, an empirically derived histrionic personality disorder (PD), and attempted to validate the Q-sort measure of it by locating it within the context of the current adult PDs defined by the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (*DSM–IV*; American Psychiatric Association, 1994). In other words, our first pass at construct validity attempted to locate this new construct and measure within the nomological network defined by the current diagnostic system. The investigators had four measures of current *DSM–IV* PD diagnosis at their disposal, but for our purposes we focus on two: clinicians' 7-point global ratings of the extent to which the patient met criteria for each of the 10 PDs listed in the *DSM–IV*, and a numerical count of the number of Axis II symptoms clinicians rated as present for each disorder using current diagnostic criteria randomly ordered (to avoid halo effects).

To quantify the validity of the new measure of this empirically derived construct of histrionic PD of adolescence, we attempted to predict the pattern of correlations between the new measure and the current 10 PD diagnoses. Although the *DSM–IV* includes a histrionic diagnosis, both the item content of the empirically derived histrionic prototype and the item content of an empirically derived histrionic prototype in an adult sample using a similar method suggested that patients high on this dimension have a number of features currently diagnosed not only as histrionic but also as borderline and, secondarily, as dependent. (This likely accounts for the substantial diagnostic overlap among these three diagnoses seen in clinical practice; that is, the current diagnoses may not adequately "carve nature at its joints"; Westen & Shedler, 1999.)

Thus, with respect to convergent validity, we predicted, in descending order, correlations between the adolescent histrionic diagnosis and current histrionic, borderline, and dependent features. With respect to discriminant validity, patients with adolescent histrionic PD appear to be extraverted and hence should be quite low on schizoid, schizotypal, and avoidant personality characteristics as defined using adult *DSM–IV* criteria. In addition, their impressionistic cognitive style should produce a negative correlation with the more detail-oriented obsessive–compulsive PD. Finally, we predicted correlations near zero with antisocial, narcissistic, and paranoid ratings, reflecting our belief that we had isolated a histrionic construct independent of the other "Cluster B" *DSM–IV* PDs and of paranoid PD, which should show no particular relation to histrionic.

Table 1
Predicted Correlations Between New Construct (Histrionic Personality Disorder of Adolescence) and Dimensional Diagnosis of 10 Axis II Disorders, Raw λs, and Integer Values of λ

1	2	3	4
Diagnoses involved in prediction	Predicted correlations with diagnoses	Demeaned correlations: Raw λs	Integer values of Raw λs
Histrionic	.60	.70	7
Borderline	.30	.40	4
Dependent	.10	.20	2
Antisocial	.00	.10	1
Narcissistic	.00	.10	1
Paranoid	−.10	.00	0
Obsessive–compulsive	−.40	−.30	−3
Avoidant	−.50	−.40	−4
Schizoid	−.50	−.40	−4
Schizotypal	−.50	−.40	−4
M	−.10	.00	0

The empirically derived portrait of the adolescent histrionic patient thus led us to predict the pattern of correlations shown in Table 1 (column 2), which were converted into raw lambda weights (column 3) and then into integer-valued lambda weights (column 4). Raw weights (λs) are constructed by subtracting the mean predicted value from each predicted value so that the sum of the "demeaned" predicted values is zero. It is convenient, but not required, to convert raw lambda weights to integer values before computing contrasts and associated quantities. The resulting convergent-discriminant validity array represents our underlying theory of the construct and measure.[1]

Columns 2 and 3 of Table 2 provide the actual (observed) correlations between histrionic scores using this new measure and (a) global PD ratings and (b) number of symptoms (using the current *DSM–IV* system), respectively. Because calculations involving these *r*s require that we transform all *r*s by means of the Fisher Z_r transformation,

$$Z_r = \tfrac{1}{2} \log_e \left(\frac{1 + r}{1 - r}\right),$$

columns 4 and 5 show these transformations for columns 2 and 3, respectively. Fisher's Z_r transformed correlations are virtually identical to nontransformed *r*s, at values from .00 to .25, and within 10% of *r*s from .25 to .50, so that the transformed and nontransformed values here (and in most cases) are very similar.

In this example, although we expected the numerical count (number of symptoms of each disorder) to be a more reliable measure than global ratings (because it includes multiple items rather than a single rating, and because it presented clinicians with the current criteria rather than expecting them to work from memory), we did not expect the patterns of correlations with our new measure to differ across the two measures of the criterion variables (current *DSM–IV* diagnoses). Thus, we made the same predictions for both measures, and hypothesized no (systematic) method variance. As we will see, in other cases, specification of method variance (e.g., self-report vs. behavioral observation) can be an important part of construct validation, and can help determine whether two putative measures of the same construct are really measuring the same thing.

The results suggest that we do, in fact, have a good theory of the relation between our new construct and measure of histrionic PD of adolescence and the current psychiatric nomenclature. Appendix A presents the results of the contrast analysis of the data of Tables 1 and 2. Information required for the computations is provided along with effect size estimates, confidence intervals, and significance levels. Appendix B shows how we arrived at these results using this data set, and provides an illustrative example of how these metrics can be used in practice.

In this example, $r_{alerting\text{-}CV}$ exceeded .90 for single-item ratings, number of symptoms, and both sets of criteria combined. In other words, the pattern of predicted correlations strongly matched the pattern of observed correlations. The coefficients for $r_{contrast\text{-}CV}$ were .715, .801, and .888, respectively, for single-item ratings, number of symptoms, and both sets of criteria combined, with exceedingly small corresponding *p*-values associated with all three values of *r*. Once again, the data show substantial correspondence between our theory of the construct and its empirical correlates. The magnitude and meaning of these *r*s should be interpreted just as other *r*s in construct validation are interpreted (e.g., as if we had simply correlated our measure with one other measure with which we expected it to correlate), and suggest that we understood the construct very well. The *p*-values suggest that our highly specific, one degree of freedom prediction about the magnitude of correlations could not likely have been obtained by chance.

The findings also demonstrate, as expected, that multiple-item criteria (number of symptoms of each disorder) performed better than single-item criteria. Combining the two sets of criterion

[1] Making predictions using lambda weights is a relatively straightforward process, in which an investigator armed with some knowledge of the construct predicts a pattern of results, and then performs simple addition and subtraction, as in this example, to convert them to integer values. What the investigator strives to optimize is not the absolute magnitude of the predicted correlations (e.g., guessing that a measure will correlate $r = .62$ with another measure) but the *relative* magnitude of a series of correlations. Researchers interested in testing the consensual nature of the predictions could, if they chose, have multiple knowledgeable raters each propose lambda weights and measure interrater reliability.

Table 2
Correlations and Z_r Transformed Correlations Between Adolescent Histrionic Personality Disorder Scores (New Measure) and Adult Personality Disorder Scores Measured by Ratings and by Number of Symptoms

1	2	3	4	5
Diagnoses involved in prediction	Ratings (r)	No. of symptoms (r)	Ratings (Z_r)	No. of symptoms (Z_r)
Histrionic	.55	.64	.62	.76
Borderline	.51	.39	.56	.41
Dependent	.20	.38	.20	.40
Antisocial	−.06	.11	−.06	.11
Narcissistic	.10	.19	.10	.19
Paranoid	−.04	.10	−.04	.10
Obsessive–compulsive	−.23	−.12	−.23	−.12
Avoidant	−.20	−.07	−.20	−.07
Schizoid	−.15	−.13	−.15	−.13
Schizotypal	−.02	−.04	−.02	−.04

variables yielded even higher estimates of validity for $r_{contrast\text{-}CV}$. Thus, the two sets of criteria provided slightly different, and complementary, sources of information about construct validity—that is, inclusion of two measures of each of the *DSM–IV* PDs maximized reliability of measurement. (The degree of similarity between the two sets of criteria is indexed by the correlation between them computed on columns 4 and 5 of Table 2, $r = .937$, or on columns 2 and 3, $r = .941$.)

To summarize our two simple methods for quantifying construct validity: To compute $r_{alerting\text{-}CV}$, we simply correlate the hypothesized pattern of correlations between our measure under investigation and various criterion variables, expressed as a pattern of lambda weights arrayed in one column and the observed pattern of correlations arrayed in a second column. The resulting correlation coefficient provides an index of the extent to which our underlying theory accurately predicts the relative magnitude of the observed pattern of correlations. It does not, however, take into consideration variables such as sample size and the size of the intercorrelations of the criterion variables, and it does not readily produce a meaningful *p*-value. That is the advantage of $r_{contrast\text{-}CV}$, which also provides data on the degree of fit between expected and observed values but takes into account both (a) the median intercorrelations among the criterion variables and (b) the average (squared) correlation between the measure being validated and the criterion variables. To compute $r_{contrast\text{-}CV}$, we test the relation between our lambda weights and the observed pattern of correlations using a simple *t* statistic, and compute $r_{contrast}$ from the associated *t*-value. (The coefficient *r* is readily obtained from knowing *t* and its degree of freedom, because *r* is a simple function of significance test and sample size.)

In practice, $r_{alerting\text{-}CV}$ tends to be larger than $r_{contrast\text{-}CV}$, but that need not be so. We recommend using both $r_{contrast\text{-}CV}$ and $r_{alerting\text{-}CV}$. In its present application, $r_{contrast\text{-}CV}$ is more computationally tied to significance tests and to the construction of confidence intervals. However, $r_{alerting\text{-}CV}$ not only provides an intuitively interpretable effect size correlation, but it also helps prevent us from dismissing a nonsignificant $r_{contrast\text{-}CV}$ as unimportant merely because it is nonsignificant (i.e., sample size is too small) or because it does not strike the investigator as very large in magnitude.

The Impact of Larger and Smaller Correlations on the Two Measures of Construct Validity

Although both $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$ are informative, they respond differently to different levels of elevation of the profile of obtained correlations. Table 3 shows the effects on $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$ of moving all the obtained correlations found in column 3, which range from .76 to −.13, closer to .00. To illustrate the relative impact on $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$, we simply divided the original Z_r transformations of the obtained correlations by 4, with a resulting range of only .19 to −.03 (column 4). Column 5 shows an even greater shrinkage toward .00 of the Z_rs of column 3 by dividing them by 10, with a resulting range of only .08 to −.01. This illustration demonstrates what happens when the investigator correctly predicts the rank order of the observed correlations but the obtained correlations are weaker than those in the example provided previously of histrionic PD of adolescence.

Because dividing a variable by a constant has no effect on its correlation with another variable, our dividing the Z_rs of column 3 by 4 or by 10 had no effect whatsoever on $r_{alerting\text{-}CV}$, which remained at its original level of .96. However, the effect on $r_{contrast\text{-}CV}$ of shrinking the Z_rs of column 3 closer toward .00 was dramatic. When the Z_rs of column 3 were reduced by a factor of 4, $r_{contrast\text{-}CV}$ dropped from .80 to .26. When the shrinkage was by a factor of 10, $r_{contrast\text{-}CV}$ dropped from .80 to .10. Table 3 shows that, as expected, as $r_{contrast\text{-}CV}$ decreases, with no change in sample size, the *p*-values associated with $Z_{contrast}$ and with the chi-square test of heterogeneity become less and less significant. These examples show the advantages of having two metrics of construct validity, one that indicates whether the investigator understands the relation between the new measure and a set of criterion variables that may be weakly related to the construct but whose relative relation to the construct and measure provides meaningful data on its validity, and the other of which is sensitive not only to whether the investigator correctly predicts the relative magnitude of the obtained correlations but also whether the correlations explain much of the variance in the new measure.

A point worth noting here is that neither $r_{alerting\text{-}CV}$ nor $r_{contrast\text{-}CV}$ provide an index of the absolute deviation of a predicted correlation (or its Z_r) from an obtained correlation (or its Z_r)—that is, of the extent to which the investigator has accurately predicted the absolute magnitudes of the observed correlations. The last two rows of Table 3 show two related indices of typical distance (D) between predicted and obtained values of Z_r: $\Sigma|D|$ and $\sqrt{\Sigma D^2/k}$ that might be useful if a researcher were interested in absolute magnitude of differences. For the present example, both these indices seem to track inversely the magnitude of $r_{contrast\text{-}CV}$ as we would hope. These and other distance metrics may prove useful in future thinking about construct validity, but they are beyond the scope of the present article. Indeed, such distance metrics of construct validation have seldom, if ever, been used, because we are seldom in the position of making such mathematically precise predictions in psychology.[2]

[2] They raise intriguing questions, however, about the circumstances under which construct validity is established by demonstrating a theoretically coherent pattern of correlations (as has traditionally been the case, and in most cases is likely the appropriate way to assess it) versus situations in which the burden of proof is higher on the investigator to specify absolute levels of association.

Table 3
Effects on $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$ of Decreasing Absolute Values of Obtained Correlations

1	2	3	4	5
Diagnoses	Predicted Z_r[a]	Original Z_r[b]	Original $Z_r/4$	Original $Z_r/10$
Histrionic	.69	.76	.19	.08
Borderline	.31	.41	.10	.04
Dependent	.10	.40	.10	.04
Antisocial	.00	.11	.03	.01
Narcissistic	.00	.19	.05	.02
Paranoid	−.10	.10	.02	.01
Obsessive–compulsive	−.42	−.12	−.03	−.01
Avoidant	−.55	−.07	−.02	−.01
Schizoid	−.55	−.13	−.03	−.01
Schizotypal	−.55	−.04	−.01	−.00
Correlations and significance tests				
$r_{alerting\text{-}CV}$		.96	.96	.96
$r_{contrast\text{-}CV}$		.80	.26	.10
$Z_{contrast}$		16.42	4.21	1.69
p of $Z_{contrast}$		$7/10^{61}$	.000013	.045
$\chi^2(9, N = 266)$		291.37	19.08	3.16
p of $\chi^2(9, N = 266)$		$2/10^{57}$	.025	.96
Differences (D) as indices of inaccuracy				
$\Sigma\|D\|/k$		.27	.29	.31
$\sqrt{\Sigma D^2/k}$		.31	.36	.39

Note. All values of Z_r are rounded to two decimal places.
[a] From Table 1, column 2. [b] From Table 2, column 5.

Discussion

Contrast analysis provides a relatively simple, readily applied method of quantifying construct validity. By specifying a set of hypotheses in advance, in the form of a predicted pattern of convergent-discriminant validity coefficients, researchers can quantify the extent to which a measure's association with other measures matches their theoretical understanding of the construct. Rather than relying on an intuitive assessment of the goodness of fit between the observed and expected pattern of correlations, this procedure allows a precise estimate.

Advantages of Quantifying Construct Validity Using Contrast Analysis

One of the advantages of this method is that, by requiring researchers to specify their predictions in the form of contrast weights, it encourages more careful thinking about ways the data might suggest refinements in the construct as well as the measure. A good illustration is the finding across a number of subfields of psychology—for example, in studies of memory, personality, emotion, motivation, psychopathology, and attitudes—of a distinction between explicit and implicit processes (see Greenwald & Banaji, 1995; Westen, 1998; Wilson, Lindsey, & Schooler, 2000). For example, research on motivation finds that self-report and projective (Thematic Apperception Test) measures of motives such as achievement tend to correlate minimally with one another, but that both predict relevant criterion variables (McClelland, Koestner, & Weinberger, 1989). In general, self-reported (explicit) motives tend to predict behavior when people are focusing their conscious attention on their goals, whereas projective or narrative-based (implicit) motives tend to predict behavior when people are behaving more habitually. Similar findings emerge in studies using priming procedures to activate implicit motives (Bargh, 1997).

A psychologist attempting to validate a new self-report measure of achievement motivation is thus likely to specify a different set of contrast weights when predicting behavior in situations designed to activate explicit achievement motivation (e.g., telling participants that they are about to take an intelligence test, and that they should try to do their very best) versus those in which achievement motives are activated implicitly (e.g., by having participants unscramble words related to achievement, such as *success* and *compete*; Bargh, 1997). Whereas a more impressionistic assessment of the multitrait–multimethod matrix might lead to the presumption that differences in magnitude of correlations reflect method variance, the requirement of contrast analysis to specify hypotheses in advance is likely to compel the researcher to think more carefully about whether the differences lie in method of assessment or in the fact that implicit and explicit achievement motivation may be distinct constructs, describing functionally and neuroanatomically different motivational systems.

A similar situation would likely occur with self-report and clinician-report measures of histrionic PD. For example, suppose clinician-report but not self-report measures of histrionic PD were to correlate with informant descriptions of patients' emotional expressions as shallow, dramatic, and superficial. This might suggest an important distinction between two constructs readily confounded by viewing the differences in terms of method variance: (a) histrionic personality patterns (assessed by observing patients' behavior) and (b) patients' explicit beliefs about their personalities (e.g., the extent to which they believe their own ways of expressing emotion are shallow and dramatic).

Caveats and Limitations

The simple metrics we are proposing are not a panacea, and we should be clear about what they can and cannot do. First, the approach to quantifying validity described in this article is not a replacement for theoretically informed judgment. It should serve as a guide to theoretically informed judgment. Correlating a self-report measure of achievement motivation with three similar self-report measures of the same construct may yield sizable validity coefficients but provide much weaker evidence for validity than a study demonstrating less robust but nonetheless impressive correlations between the measure and several behavioral indices of achievement motivation that do not share method variance (or, more precisely, that do not all involve participants' self-perceptions of their motives). This is where meta-analytic methods might prove particularly useful, by weighting studies based on variables such as sample size, number of correlations on which the validity coefficient is based (because predicting the relative magnitude of a larger set of correlations contextualizes the construct and measure within a broader nomological net), and the extent to which the convergent-discriminant validity array includes highly divergent methods. Meta-analytic procedures could also be used to detect moderator variables, such as method factors (e.g., self-report vs. behavioral criteria) or differences in samples (e.g., clinical vs. nonclinical, males vs. females), that could influence the extent to which a measure could be considered valid for different populations.[3] An advantage of using quantifiable metrics such as those described here is that they permit aggregation of data across studies.

Second, and related, consumers of research using these metrics need to be informed consumers. A researcher who specifies a series of lambda weights that predicts that a measure should correlate .2 with a measure of the same construct (with which it should correlate highly) using a different method, .6 with a measure of an unrelated trait with which it shares method variance, and .1 with a measure of an unrelated measure with which it does not share method variance could produce a high validity coefficient. In this case, however, what the researcher has done is to demonstrate that method variance swamps true variance—and indeed, the metrics described in this article can be used either to help validate a measure or demonstrate the effect size of method variance.

Third, although we have described $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$ as methods of estimating construct validity, that is not, strictly speaking, accurate. Modern accounts of construct validity include content validity and criterion validity, and although these metrics can summarize the latter, they are not intended to index the extent to which a psychologist has adequately sampled the domain in constructing a measure (content validity). However, a measure with poor content validity is unlikely to correlate highly with measures of related constructs that have more comprehensive item content representative of the domains that constitute the construct.

Finally, $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$ do not summarize an *entire* multitrait-multimethod matrix, which would provide the answer to a different question than researchers usually want to ask. Rather, they summarize the row or column of that matrix that describes the correlation between a single measure of interest (whose validity the investigator is attempting to assess) and the measures the investigator is using to try to establish its convergent and discriminant validity. Suppose, for example, an investigator is assessing the validity of a self-report measure of depression by correlating it with a self-report measure of dependency and measures of the same two constructs (depression and dependency) from behavioral observation. Table 4 describes the observed pattern of correlations. As the table shows, the observed correlations would yield a high estimate of construct validity if the investigator predicted a large correlation between the two measures of depression (monotrait, heteromethod), a medium correlation between self-reported depression and self-reported dependency (heterotrait, monomethod), and a very small correlation between self-reported depression and observed dependency (heterotrait, heteromethod). All of the information relevant to the test of the construct validity of the new measure lies on the first row of the table.

This would be a theoretically justifiable set of predictions, recognizing the likely influence of both trait and method variance, and accounting for the latter in the lambda weights. The resulting $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$ would not (and should not) reflect the fact that the self-report dependency measure did not fare as well (columns C and D), because the construct validity of that measure was not the question. Once again, this is the advantage of asking a focused question with one degree of freedom, rather than asking multiple questions simultaneously and obtaining a global index of goodness of fit that describes the entire pattern of data. There may be important uses of a metric that summarizes an entire multitrait-multimethod matrix, but our goal here is to describe metrics that summarize one column or row at a time, because this is typically the goal of the researcher studying the construct validity of a single measure.

In the present approach to quantifying construct validity, knowledge about the reliability and validity of the criterion variables is crucial for framing good hypotheses (embodied in the lambda weights), but the reliability and validity of the other measures is neither the question the researcher is trying to ask nor the answer the metrics provide. Similarly, our two metrics do not answer the question, "How substantial is method variance in assessing this construct?" (although one could certainly ask that question by focusing on a different row or column of data and making specific predictions about the magnitude of effects of method across traits). Rather, they answer the question, "How valid is this measure of this construct?" in the context of the researcher's knowledge of the likely impact of method variance, which influences the size of the predicted correlations.

Conclusions

Construct validity is a central concept in personality, clinical, educational, industrial–organizational, and other areas of psychology. Intuitive appraisals of the extent to which a construct and measure can be located within a nomological network, like intui-

[3] The ability to build moderators such as gender differences into a single metric of construct validity is beyond the scope of the method we are proposing here, which may call, instead, for the investigator to have theoretically informed "hunches" about potential moderators, such as gender, which the investigator then tests by examining the data separately by group. This is an instance in which SEM may provide useful additional information.

Table 4
Multitrait–Multimethod Matrix: Depression and Dependency Measured by Self-Report and Behavioral Observation

	Depression		Dependency	
Measures	Self-report A	Behavioral observation B	Self-report C	Behavioral observation D
A. Depression (self-report)	—	.50	.30	.10
B. Depression (behavioral observation)		—	.20	.30
C. Dependency (self-report)			—	.30
D. Dependency (behavioral observation)				—

tive appraisals of the magnitude of association between any two or more variables in a qualitative review of a literature, have their strengths and weaknesses, and depend heavily on the skill and motivations of the reviewer. Just as meta-analytic techniques can be useful in providing more precise estimates of the strength of associations among variables across studies, metrics that summarize the magnitude of association between a variable and criterion variables with which it should converge, diverge, or show little or no association should prove useful in providing more precise estimates of the extent to which we understand our constructs and measures, within and across studies.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Macmillan.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer Jr. (Ed.), *Advances in social cognition* (Vol. 10, pp. 1–61). Mahwah, NJ: Erlbaum.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Cudeck, R. (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics, 13,* 131–147.

Greenwald, A. G., & Banaji, M. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait–multimethod matrix and the representative design of experiments. *Psychological Bulletin, 100,* 257–269.

Kenny, D. A. (1995). The multitrait–multimethod matrix: Design, analysis, and conceptual issues. In P. Shrout & S. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 111–124). Mahwah, NJ: Erlbaum.

McClelland, D., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96,* 690–702.

McCrae, R., & Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175.

Meyer, G., Finn, S., Eyde, L., Kay, G. G., Moreland, K., Dies, R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Reichardt, C., & Coleman, S. C. (1995). The criteria for convergent and discriminant validity in a multitrait–multimethod matrix. *Multivariate Behavioral Research, 30,* 513–538.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* New York: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Shrout, P., & Fiske, S. (1995). *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske.* Hillsdale, NJ: Erlbaum.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251.

Westen, D. (1998). The scientific legacy of Sigmund Freud. Toward a psychodynamically informed psychological science. *Psychological Bulletin, 124,* 333–371.

Westen, D., & Chang, C. M. (2000). Adolescent personality pathology: A review. *Adolescent Psychiatry, 65,* 65–100.

Westen, D., & Shedler, J. (1999). Revising and assessing Axis II, Part 2: Toward an empirically based and clinically useful classification of personality disorders. *American Journal of Psychiatry, 156,* 273–285.

Westen, D., Shedler, J., Durrett, C., Glass, S., & Martens, A. (in press). Personality diagnoses in adolescence: *DSM-IV* Axis II diagnoses and an empirically derived alternative. *American Journal of Psychiatry*.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101–126.

Wothke, W. (1995). Covariance components analysis of the multitrait–multimethod matrix. In P. Shrout & S. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 125–144). Mahwah, NJ: Erlbaum.

Appendix A

Quantifying Construct Validity for Three Related Measures of Adolescent Histrionic Personality Disorder

Here we briefly describe the meaning of each term in Table A1 below and show how it is computed.

The quantity N is simply the total number of sampling units (clinicians or their patients in this case). The quantity k is the number of predicted correlations. In this example there are 10 such correlations for the ratings measure, 10 for the number of symptoms measure, and 20 for the combined measure that uses all the predicted values (λs) and all the obtained correlations for both measures. $\Sigma\lambda^2$ is the sum of the squared contrast weights (λs); $\Sigma\lambda Z_r$, is the sum of the products of each λ multiplied by its associated Z transformed $r(Z_r)$.

The quantity $\overline{r^2}$ is the average of the squared values of the 10 rs shown in Columns 2 and 3 of Table 2; r_x is the median intercorrelation of the "predictor" variables, that is, the median of $(10 \times 9)/2 = 45$ intercorrelations, .113 in this example.

The quantity f is defined as $(1 - r_x)/2(1 - \overline{r^2})$, which must be less than or equal to 1.00. If computation yields $f > 1$, f should be set at 1.00.

The quantity h is defined as $(1 - f\overline{r^2})/(1 - \overline{r^2})$.

The test of significance of the contrast, $Z_{contrast}$, is computed from Equation A1 as follows:

$$Z_{contrast} = \sum \lambda Z_r \sqrt{\frac{(N - 3)}{\sum \lambda^2 (1 - r_x) h}}. \tag{A1}$$

We can get an intuitive feel for this equation if we rewrite it in prose as Significance test = remarkableness of Size of Contrast × Size of Study, where

$$\text{significance test} = Z_{contrast},$$

$$\text{size of study} = \sqrt{N - 3}, \text{ and}$$

$$\text{remarkableness of size of contrast} = \frac{\sum \lambda Z_r}{\sqrt{(\sum \lambda^2)(1 - r_x)h}}.$$

The numerator of this index, $\Sigma\lambda Z_r$, is zero when the null hypothesis is true, so the larger $\Sigma\lambda Z_r$ the more remarkable the contrast size, other things being equal.

In the denominator, the first term is $\sqrt{\Sigma\lambda^2}$, which can be viewed as a "metric adjustor." That is, because we can use any set of numbers adding to zero as our contrast weights, we have to adjust for the particular metric chosen. The quantity $\Sigma\lambda Z_r$ for λs of −3, −1, +1, +3, and corresponding Z_rs of .1, .2, .3, and .6 = 1.6. However, if we substituted λs of −30, −10, +10, +30, a perfectly legitimate set of λs, our $\Sigma\lambda Z_r$ would jump to 16, larger by a factor of 10. However, adjusting our $\Sigma\lambda Z_r$ of 1.6 by the denominator term $\sqrt{\Sigma\lambda^2} = \sqrt{20}$, yields $\Sigma\lambda Z_r/\sqrt{\Sigma\lambda^2} = 1.6/\sqrt{20} = .3578$. Similarly, adjusting our $\Sigma\lambda Z_r$ of 16 by $\sqrt{\Sigma\lambda^2} = \sqrt{2000}$ yields $\Sigma\lambda Z_r/\sqrt{\Sigma\lambda^2} = 16/\sqrt{2000} = .3578$, the very same value after the appropriate adjustment. In other words, the term $\sqrt{\Sigma\lambda^2}$ keeps us from getting any unwarranted benefit from simply using larger absolute value contrast weights. For any given value of $\Sigma\lambda Z_r$, that value is more remarkable (further from the null value), the smaller the quantity $\sqrt{\Sigma\lambda^2}$.

The remaining two terms of the remarkableness index are $(1 - r_x)$ and h. Both of these are a function of two more fundamental values, r_x and $\overline{r^2}$.

Table A1
Quantities Obtained for Three Measures

Quantities	Ratings	No. of symptoms	Combined
N	266	266	266
k	10	10	20
$\Sigma\lambda^2$	128	128	256
$\Sigma\lambda Z_r$	9.19	9.38	18.57
$\overline{r^2}$	.0734	.0802	.0768
r_x	.113	.366	.168
f	.4786	.3446	.4506
h	1.0413	1.0571	1.0457
$Z_{contrast}$[a]	13.71	16.42	20.18
p	$5/10^{43}$	$7/10^{61}$	$8/10^{91}$
$t_{contrast}$[b]	16.59	21.72	31.26
$r_{contrast\text{-}CV}$[c]	.715	.801	.888
Standard error of $\Sigma\lambda Z_r$[d]	.6705	.5711	.9203
95% confidence intervals for $\Sigma\lambda Z_r$[d]			
from:	7.88	8.26	16.77
to:	10.50	10.50	20.37
$r_{alerting\text{-}CV}$	.9036	.9622	.9219
$r^2_{alerting\text{-}CV}$	.8165	.9258	.8498
$\chi^2(k - 1)$[e]	230.13	291.37	479.16
$Z_{contrast} = [r^2_{alerting\text{-}CV} \times \chi^2\ (k - 1)]^{1/2}$	13.71	16.42	20.18
95% confidence intervals for $r_{contrast\text{-}CV}$			
from:	.651	.753	.860
to:	.769	.841	.911

[a] Equation 6 from Meng, Rosenthal, & Rubin, 1992. [b] From p values associated with $Z_{contrast}$. [c] Equation 2.3 from Rosenthal, Rosnow, & Rubin, 2000. [d] Equation 7 from Meng et al., 1992. [e] Equation 5 from Meng et al., 1992.

(*Appendixes continue*)

As noted earlier, r_x is the median intercorrelation among the "predictor" variables, and the larger it is, the more remarkable is the size of the contrast, because any given level of $\Sigma\lambda Z_r$ is harder to achieve when the predictors are more like each other.

As noted earlier, the quantity $\overline{r^2}$ is the average squared correlation between the measure being validated and the criterion variables. As this quantity, $\overline{r^2}$, gets either quite small (moving toward .00) or quite large (moving toward 1.00), the more remarkable is the size of the contrast, because any given level of $\Sigma\lambda Z_r$ is harder to achieve when the predictors show less variability in their correlations with the measure being validated. In the extreme case, of course, if the median $\overline{r^2} = .00$ or 1.00, there can be no variability whatsoever, and so our significance test Z and its associated effect size r must both be .00.

The value p is simply the significance level associated with $Z_{contrast}$. The value $t_{contrast}$ is the exact value of t associated with the exact value of p. The value of $r_{contrast}$ is obtained from $t_{contrast}$ as follows in Equation A2:

$$r_{contrast} = \sqrt{\frac{t^2_{contrast}}{t^2_{contrast} + df}}. \tag{A2}$$

The standard error (SE) of the contrast, $\Sigma\lambda Z_r$, is given by Equation A3, as follows:

$$SE = \sqrt{\frac{\Sigma\lambda^2(1 - r_x)h}{N - 3}}, \tag{A3}$$

and the 95% confidence interval is given by 1.96 SE added to $\Sigma\lambda Z_r$ for the upper limit and 1.96 SE subtracted from $\Sigma\lambda Z_r$ for the lower limit.

The quantity $r_{alerting}$ is the correlation between the contrast weights and their associated obtained correlations after Z_r transformation. The quantity $r^2_{alerting}$ is simply the squared value of $r_{alerting}$.

The value $\chi^2(k - 1)$ is the χ^2, with $k - 1$ degrees of freedom, testing for the heterogeneity of the set of "predictor" variables; 10 in the present example for the ratings measure, 10 for the number of symptoms measure, and 20 for the combined measure. Equation A4 shows how to obtain $\chi^2(k - 1)$:

$$\chi^2(k - 1) = \frac{(N - 3)\Sigma(Z_r - \overline{Z}_r)^2}{(1 - r_x)h}, \tag{A4}$$

where $\overline{Z}_r$ is the mean of the Z_rs and all other terms are as defined earlier.

An alternative procedure for computing $Z_{contrast}$ is available simply by multiplying $r^2_{alerting}$ by $\chi^2(k - 1)$ yielding a $\chi^2_{contrast}$ of $df = 1$.

The square root of this $\chi^2_{contrast}$ (1) = $Z_{contrast}$, that is,

$$Z_{contrast} = \sqrt{r^2_{alerting} \times \chi^2(k - 1)}\,. \tag{A5}$$

Finally, we can also compute a useful, approximate 95% confidence interval for $r_{contrast}$ by transforming $r_{contrast}$ to its associated Z_r and obtaining the upper and lower 95% limits from $Z_r \pm 1.96/\sqrt{N - 3}$. As a final step in computing these confidence intervals we transform the upper and lower limit Z_rs to their corresponding rs.

Appendix B

Calculating $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$

We illustrate the computations for the ratings data of Tables 2 and 3. The value N is simply the number of sampling units employed in the analysis—266 in this example. The value k is the number of predicted correlations—10 in this example. $\Sigma\lambda^2$ adds the squared λs of Table 1, that is,

$$(7)^2 + (4)^2 + \ldots + (-4)^2 + (-4)^2 = 128;$$

$$\sum\lambda Z_r = 7(.62) + 4(.56) + \ldots + (-4)(-.15) + (-4)(-.02) = 9.19;$$

$$\overline{r^2} = [(.55)^2 + (.51)^2 + \ldots + (-.15)^2 + (-.02)^2]/10 = .0734.$$

The quantity $r_x = .113$ is the median of the 45 intercorrelations among the 10 "predictors." The quantity f is computed from $(1 - r_x)/2(1 - \overline{r^2})$, which, in our example, yields $(1 - .113)/2(1 - .0734) = .4786$. The quantity h is defined as

$$(1 - f\overline{r^2})/(1 - \overline{r^2}) = 1 - .4786(.0734)/(1 - .0734) = 1.0413.$$

We find $Z_{contrast}$ from Equation A1 (see Appendix A) as

$$Z_{contrast} = 9.19\sqrt{\frac{266 - 3}{128(1 - .113)1.0413}} = 13.71,$$

which has an associated p of $5/10^{43}$, a quantity available from computers and many calculators. Again, using these calculators or computers, it is easy to obtain the $t_{contrast}$ associated with this p entering only the df for $t(N - 2)$; $t_{contrast} = 16.59$ in this example.

We obtain $r_{contrast}$ as shown in Equation A2 (see Appendix A), in this case finding

$$r_{contrast} = \sqrt{\frac{(16.59)^2}{(16.59)^2 + (266 - 2)}} = .715.$$

We find the standard error of the contrast $\Sigma\lambda Z_r$ from Equation A3 (see Appendix A) as

$$SE = \sqrt{\frac{128(1 - .113)1.0413}{266 - 3}} = .6705,$$

and because the 95% confidence interval is given by $\Sigma\lambda Z_r \pm 1.96\ SE$, the interval extends from 7.88 to 10.50.

We compute $r_{alerting}$ directly as the correlation between our contrast weights and our Z_r transformed correlations and find

$$r_{alerting} = .9036 \text{ and } r^2_{alerting} = .8165.$$

From Equation A4 (see Appendix A) we compute the $\chi^2(k - 1)$ as follows:

$$\chi^2(k - 1) = \frac{\begin{array}{c}(266 - 3)\ [(.62 - .078)^2 + (.56 - .078)^2 \\ + \ldots + (-.15 - .078)^2 + (-.02 - .078)^2]\end{array}}{(1 - .113)1.0413} = 230.13.$$

We can also compute $Z_{contrast}$ from Equation A5 (see Appendix A) as follows:

$$Z_{contrast} = \sqrt{.8165(230.13)} = 13.71.$$

Finally, we can compute the 95% confidence interval for $r_{contrast}$ beginning by transforming our $r_{contrast}$ of .715 to its Fisher Z_r transform of .897, and finding the lower and upper limits for Z_r from $Z_r \pm 1.96/\sqrt{266 - 3} = .897 \pm .1209$ as .776 to 1.018 in the Z_r transform. Converting these lower and upper limits back into the metric of r yields lower and upper 95% confidence limits of .651 to .769.

Received January 18, 2002
Revision received August 13, 2002
Accepted August 22, 2002 ■