# 3

# ON CREATING AND USING SHORT FORMS OF SCALES IN SECONDARY RESEARCH

KEITH F. WIDAMAN, TODD D. LITTLE, KRISTOPHER J. PREACHER, AND GITA M. SAWALANI

Short forms are so-named because they consist of shortened versions of original scales. Creating and using such measures is something of a cottage industry for many practicing scientists. A PsycINFO search conducted in late December 2008 using *short form* and *short forms* as alternate keywords returned 5,101 references, over 4,200 of which were articles in refereed journals. Many of these references had an applied measurement or methodological focus, describing the creation and psychometric evaluation of a short form of an extant instrument. For example, Rammstedt and John (2007) recently created a 10-item version of the 44-item Big Five Inventory (John, Donahue, & Kentle, 1991) as a way to measure the five broad dimensions of personality (i.e., Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness) in 1 minute or less. Researchers often create short forms so they can assess a larger number of constructs in a reasonably short testing session. One particular advantage is that use of short forms can ensure that measures of many potentially critical variables are obtained so the researcher has greater latitude in testing alternative hypotheses.

Regardless of any compelling reasons for using short forms, the most likely outcome of using them is that the researcher will have a measure of the construct that has poorer psychometric properties than does the original long

form. As a consequence, researchers must make careful and wise decisions about how to analyze the data and interpret the results. In this chapter, we provide an overview of issues related to reliability and validity, and provide explicit recommendations for creating and using short forms in behavioral science research, particularly as they apply to existing data sets. In this regard, we focus primarily on the typical secondary data set—an existing data set collected by other researchers to pursue their goals, yet a data set that can be mined to answer current theoretical questions. In most secondary data sets, many constructs are assessed using self-report questionnaire formats, although these data sets often include constructs assessed using other measurement methods, such as observer ratings, informant reports, and objective tests.

Researchers who use secondary data often must make decisions about which scales to use, whether to construct new scales from existing sets of items, and how to evaluate scales once they are assembled from existing data, among other concerns. This task can be daunting, particularly when considering whether one can "sell" reviewers of a manuscript on the nonstandard nature of certain measures. Therefore, our goal in this chapter is to provide applied researchers with a tutorial on developing and evaluating short forms derived from secondary data.

## FUNDAMENTAL ISSUES IN PSYCHOLOGICAL MEASUREMENT: RELIABILITY

One very important psychometric property of a measure is its reliability. *Reliability* refers to precision of measurement—or how accurately determined an individual's score is on a measure or scale. In classical test theory (e.g., Lord & Novick, 1968), reliability is defined as the ratio of true score variance over total scale variance. As the ratio of true score variance over total variance increases, the relative proportion of error variance decreases, so precision of measurement is positively related to reliability and inversely related to error variance in a scale.

Precision of measurement is often presented in terms of the standard error of measurement for an instrument. The *standard error of measurement* is the standard deviation of a sampling distribution centered at a person's true score and estimates the standard deviation of scores the person would obtain on an infinite number of administrations of the instrument, other things being equal. If reliability of scale X is denoted $r_{XX}$, the standard error of measurement is calculated as $s_X\sqrt{1 - r_{XX}}$, where $s_X$ is the standard deviation of scale X. Reliability is a key component in the formula for the standard error of measurement: As reliability increases, the standard error of measurement decreases, and precision of measurement improves.

## Types of Reliability Coefficient

Reliability involves the extent to which individual differences in measured scores are consistent and reproducible. Scales and tests with adequate reliability should yield more or less the same scores across periods of time and groups. Empirically speaking, reliability can be indexed in a number of ways.

### *Internal Consistency*

Many different kinds of reliability coefficient have been developed, such as split half, internal consistency, parallel forms, and test–retest reliability (see McDonald, 1999). The most commonly used reliability coefficients are internal consistency and test–retest coefficients. Internal consistency indexes estimate reliability on the basis of associations among scale items. Coefficient alpha (Cronbach, 1951) is the most often reported internal consistency index. However, many researchers are unaware that alpha is based on the assumptions that a single factor underlies the scale and that all items are equally good indicators of the latent variable being assessed (Schmitt, 1996). That is, if an item factor analysis were performed, all loadings of the $p$ items on the single factor underlying the scale would be equal (i.e., tau equivalent). If the factor loadings are not equal for all items, coefficient omega is a more appropriate estimator of reliability (see below), and coefficient omega is always greater than or equal to coefficient alpha for a scale that is unidimensional (i.e., that is a one-factor scale; see McDonald, 1970, 1999).

Coefficient alpha can be calculated in many ways, but perhaps the easiest way is

$$\alpha = \left(\frac{p}{p-1}\right)\left(\frac{s_X^2 - \sum s_j^2}{s_X^2}\right) = \left(\frac{p}{p-1}\right)\left(1 - \frac{\sum s_j^2}{s_X^2}\right), \qquad (1)$$

where $p$ is the number of items, $s_X^2$ is the variance of total scores on scale X, and $\sum s_j^2$ refers to the summation of item variances for the $p$ items ($j = 1, \ldots, p$).

In the left section of Table 3.1, descriptive statistics are presented for a 6-item short form of the 10-item Rosenberg Self-Esteem Scale (Rosenberg, 1965; note that Items 5 and 6 are reversed scored), which are based on a sample of 6,753 participants in the 2005 Monitoring the Future survey. These data are freely available to researchers (more details are available at http://monitoringthefuture.org/).

The correlations among items, shown below the main diagonal, are moderate to large, ranging between .256 and .696, with a mean correlation of .468. Item variances are shown on the diagonal, and covariances among items are above the diagonal. To use Equation 1 to calculate coefficient alpha, we

TABLE 3.1
Six Rosenberg Self-Esteem Scale Items From the 2005 Monitoring The Future Study: Descriptive Statistics and One-Factor and Two-Factor Solutions

| | Item descriptive statistics | | | | | | One-factor model | | Two-factor model | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 | 5r | 6r | $\lambda_1$ | $\theta_j^2$ | $\lambda_1$ | $\lambda_2$ | $\theta_j^2$ |
| 1 | **1.094** | **.660** | **.495** | **.783** | **.470** | **.553** | .853 | .367 | .810[b] | .058[a] | .377 |
| 2 | .613 | **1.061** | **.569** | **.661** | **.440** | **.456** | .776 | .458 | .825[b] | −.042[a] | .420 |
| 3 | .516 | .602 | **.843** | **.525** | **.297** | **.300** | .604 | .478 | .708[b] | −.121[a] | .432 |
| 4 | .696 | .596 | .532 | **1.158** | **.526** | **.597** | .886 | .372 | .806[b] | .105[a] | .393 |
| 5r | .381 | .362 | .274 | .415 | **1.390** | **.878** | .605 | 1.024 | .094[b] | .750[a] | .734 |
| 6r | .415 | .347 | .256 | .435 | .584 | **1.626** | .670 | 1.177 | −.094[b] | 1.149[a] | .427 |
| *M* | 3.978 | 4.072 | 4.171 | 4.045 | 3.944 | 3.753 | | | | | |
| *SD* | 1.046 | 1.030 | .918 | 1.076 | 1.179 | 1.275 | | | | | |

*Note.* $N = 6{,}753$. In the Item Descriptive Statistics section, values are correlations among items (below diagonal), item variances (on diagonal in bold print), and covariances among items (above diagonal, in bold print), with the mean and standard deviation for each item. Items 5 and 6 were reversed scored (resulting in Items 5r and 6r, respectively) prior to calculating item statistics. In the "One-factor" and "Two-factor" sections, tabled values are estimates from factor analyzing covariances among items. Symbols $\lambda_1$ and $\lambda_2$ refer to Factors 1 and 2, respectively, and $\theta_j^2$, to unique factor variance. Factor variances were fixed at unity to identify and scale the estimates. Factors correlated .608 in two-factor solution.

[a]Parameters constrained to sum to zero. [b]Parameters constrained to sum to zero.

need two quantities: (a) the sum of all elements of the item variance–covariance matrix, which is equal to $s_X^2$, is 23.592, and (b) the sum of the item variances on the diagonal, which is 7.172. Given these values and the presence of $p = 6$ items, coefficient alpha is estimated as $\alpha = \left(\frac{6}{6-1}\right)\left(\frac{23.592 - 7.172}{23.592}\right) = .835$ for this six-item short form.

An interesting alternative to coefficient alpha is coefficient omega (McDonald, 1970, 1999). Coefficient omega is more appropriate for most research applications because it is unrealistic to assume that all items on a measure are equally good at tapping true score variance. In our experience, equal factor loadings for all scale items rarely, if ever, occur, so coefficient omega will, in general, be preferable to coefficient alpha. Assuming that a linear model underlies responses on each item on scale X, the linear model for item $x_{ji}$ takes the form $x_{ji} = \tau_j + \lambda_j F_i + \varepsilon_{ji}$, where $x_{ji}$ is the score of person $i$ on item $j$, $\tau_j$ is the intercept (i.e., mean) of item $j$, $\lambda_j$ is the raw score (or covariance metric) common factor loading for item $j$, $F_i$ is the score on the common factor for person $i$, and $\varepsilon_{ji}$ is the score of person $i$ on the unique factor for item $j$. On the basis of factor analysis of the item variance–covariance matrix, coefficient omega can be estimated as

$$\omega = \frac{\left(\sum \lambda_j\right)^2}{\left(\sum \lambda_j\right)^2 + \sum \theta_j^2} = 1 - \frac{\sum \theta_j^2}{s_X^2}, \qquad (2)$$

where all summations are from 1 to $p$ (the $p$ items), $\theta_j^2$ is the estimated unique variance of item $j$ (i.e., the variance of $\varepsilon_{ji}$), and other symbols are as defined previously. In the first expression in Equation 2, one must first sum the $p$ factor loadings and square this sum; the square of summed loadings estimates the variance of the scale. The denominator is the sum of the preceding value and the sum of the unique variances of the items. The ratio of these two values gives the proportion of variance in the scale that is reliable variance (i.e., coefficient omega). When a single factor underlies the scale and all factor loadings ($\lambda_j$ s) are identical, coefficients omega and alpha are identical.

The second expression in Equation 2 is simply 1.0 minus the ratio of the sum of unique factor variances, $\sum \theta_j^2$, over total scale variance, $s_X^2$. If a scale is truly a single-factor instrument (i.e., is unidimensional), both expressions for coefficient omega in Equation 2 will provide identical results. But recent work by Zinbarg and Revelle and associates (e.g., Zinbarg, Revelle, & Yovel, 2007; Zinbarg, Yovel, Revelle, & McDonald, 2006) showed that the second expression is a more appropriate estimator of coefficient omega as developed by McDonald (1970) if a scale is "lumpy" (cf. Cronbach, 1951) and thus consists of two or

more highly correlated group factors that reflect excess overlap in item content or stylistic variance that contributes to the multidimensionality.

To estimate coefficient omega, we used maximum likelihood estimation in Mplus (Muthén & Muthén, 1998–2007) to obtain a one-factor solution for the self-esteem items. To identify this model, we fixed the factor variance to 1.0 and estimated all remaining parameters. As seen in Table 3.1, the positively worded items had higher factor loadings and much lower unique variances than did the negatively worded items. Factor loadings varied considerably (range = .604–.886) and thus were not tau equivalent, implying that coefficient alpha is inappropriate for this data set. The one-factor solution had marginal levels of fit to the data, with a comparative fit index (CFI) of .881 and standardized root-mean-square residual (SRMR) of .069, so computation of coefficient omega may be suspect. For illustration, we used the first expression for coefficient omega shown in Equation 2 to compute a reliability estimate (the sum of factor loadings is 4.394, the square of this sum is 19.307, and the sum of unique variances is 3.876). This estimate of coefficient omega was .833, which is marginally lower than coefficient alpha reported above. Using the second expression in Equation 2, coefficient omega was .836, marginally higher than coefficient alpha. That the first of these two estimates of coefficient omega is lower than coefficient alpha is inconsistent with the claim that omega is greater than or equal to alpha (McDonald, 1999), an inequality that holds only if a scale is unidimensional. Thus, the first estimate of coefficient omega suggests that the one-factor solution for this data set is inappropriate.

To investigate this, we fit a freely rotatable, exploratory two-factor model to the data, again using maximum likelihood estimation and the Mplus program. To identify this model, we fixed factor variances to 1.0, allowed the factors to correlate, and constrained hyperplanar loadings on each factor to sum to zero. That is, we constrained the loadings of Items 5r and 6r on the first factor to sum to zero, and we constrained the loadings of Items 1 through 4 on the second factor to sum to zero. The results of this analysis are shown in the last three data columns of Table 3.1. The two-factor solution had quite acceptable levels of fit, with CFI of .980 and SRMR of .018. As seen in Table 3.1, the four positively worded items loaded highly on the first factor, the two negatively worded items loaded highly on the second factor, and the two factors were relatively highly correlated (.608). Notably, the unique factor variances for the two negatively worded items were greatly reduced relative to the one-factor model, and the sum of unique variances was now 2.783. Using the second expression in Equation 2, omega was 882, considerably higher than coefficient alpha.

Coefficient alpha is often touted as a lower bound estimator of scale reliability, and this sounds at first to be desirable, as one generally would not

like to overstate the reliability of a measure. However, we think the most accurate estimate of scale reliability is a more useful value. As such, we advocate reporting coefficient omega given that the assumption of tau-equivalent items is unlikely to hold. We also recommend reporting mean interitem correlations (MIC) more frequently, so that estimates of reliability across scales with varying lengths can be compared more easily.

### *Test–Retest Reliability*

The second most commonly reported index of reliability is test–retest reliability. Here, one administers the same scale at two points and calculates the Pearson product–moment correlation between the two administrations. In secondary data sets that contain longitudinal measurements, correlations between scores on a given scale across measurement occasions are test–retest correlations, and even single-item measures can be evaluated using test–retest reliability.

The strength of test–retest correlations depends on the time lag between assessments, with longer lags tending to yield lower correlations. Therefore, such estimates are not optimal as indices of measurement precision; instead, they reflect stability over time. Additionally, internal consistency and test–retest reliabilities may diverge as a function of the type of construct. A scale assessing a trait construct would ideally have high internal consistency reliability and should also exhibit high test–retest reliability. In contrast, if a scale assesses a state or mood construct that varies across time, the scale would ideally have high internal consistency reliability, whereas its test–retest reliability should be quite low, even near zero. If the test–retest reliability of a state or mood scale is quite high, the contention that the scale assessed a state construct is open to question.

## Scale Characteristics Affecting Reliability

To better understand reliability, we delve into some of the characteristics of scales that can change the reliability by which a construct is measured. These features are not exhaustive, but they are central to understanding the reliability of any measurement tool.

### *Three Key Features*

Three principal features influence the reliability of a scale: (a) the number of items in the scale, (b) the magnitude of the MIC among the items, and (c) the relative standard deviations of the items. With regard to the first of these, other things being equal, the larger the number of items on a scale, the higher the reliability of the scale. Because items composing a short form

are a subset of items in the original scale, a short form will usually have lower reliability than the original, longer form. This reduced reliability is a crucial consideration: If a short form is too short, its reliability may be so compromised that it has unacceptable levels of measurement precision—and its use in research becomes a risky endeavor. Providing bounds of reliability that are acceptable is fraught with problems. In general, short-form scale reliabilities of .80 or above are generally quite acceptable, values between .70 and .80 are adequate for research purposes, and reliabilities between .60 and .70 are at the low end of general use (cf. Nunnally & Bernstein, 1994, pp. 264–265). However, occasionally a scale with a reliability of .45 to .50 has surprisingly high correlations with outside variables, so the proof of its use is in the empirical relations it has with other variables (i.e., its validity, as discussed later).

The second scale characteristic—the magnitude of the MIC—reflects the amount of variance that is shared among items. Here, the higher the MIC, the smaller the number of items needed to achieve an acceptable level of reliability. Conversely, scales with a low MIC will require more items to achieve a comparable level of reliability. Clark and Watson (1995) argued that the MIC for scales generally should fall somewhere between .15 and .50. When measuring broad constructs like extroversion, lower MICs (in the range from .15 to .35) are expected, as one should attempt to cast a broad net and assess many, somewhat disparate, aspects of a content domain. When measuring a narrow construct like test anxiety, higher MICs (ranging from .30 to .50) should occur because narrow domains of content have items of greater similarity.

The MIC is directly related to the standardized factor loadings one would obtain if item correlations were factor analyzed. If correlations among items were fairly homogeneous and the MIC were .16, then standardized factor loadings for items would be around .40, and MICs of .25, .36, and .50 would translate into factor loadings of about .50, .60, and .70, respectively. Thus, the guideline by Clark and Watson (1995) that the MIC should fall between .15 and .50 means that standardized item factor loadings should vary between about .40 and .70, a useful benchmark when evaluating factor analyses of items comprising a short form from a secondary data set.

With regard to the third scale characteristic, differences in item variance can affect both reliability and the nature of the construct assessed by the scale. Other things being equal, items with larger variance contribute proportionally more variance to the scale than do items with smaller variance. When this occurs, individual differences on items with larger variance contribute disproportionately to individual differences on the total scale, which is shifted in the direction of these items and away from items with relatively small variance. When items with small variance assess crucial aspects of a construct, failure

to take account of differences in item variance lowers the contribution of these components to the scale's content domain.

If all items fall on the same (or a similar) scale, differences in variance across items are generally small. For example, if each item is rated on a 1-to-5 scale, then marked differences in variance across items are unlikely and would not have large differential effects on scale scores.

However, items in a short form may have rather different scales (e.g., some items might fall on 1-to-3 rating scales, others on 1-to-5 rating scales, and still others on 1-to-10 or even 0-to-100 rating scales). When scale formats vary considerably across indicators, items must be rescaled so that items rated using larger scales do not bias the summary measure too far in their direction. One alternative is to convert all indicators to $z$ scores, with $M$ of 0.0 and $SD$ of 1.0; if items have identical means and standard deviations, then differential contributions to the scale score are minimized. However, standardization should be done with care, because it is a sample-specific operation that will not generalize to a new sample. Furthermore, if existing data have a longitudinal aspect, care must be taken to transform a given item in a comparable fashion (e.g., using the same mean and standard deviation) at each time point so that scores are on a comparable metric across time. A second option is to use percent of maximum possible (POMP) scoring (see, e.g., Cohen, Cohen, Aiken, & West, 1999). Here, each item is rescaled to fall between 0 and 1, and then averaged into a scale score. For example, if an item were originally scored on a 1-to-5 scale, the researcher could subtract 1 from each person's score (so scores would now fall on a 0-to-4 scale) and then divide by 4. POMP scoring has the advantage that all item scores fall on a scale with the same potential range (i.e., 0 to 1), although items may still differ markedly in variance. Thus, POMP scoring may not be an optimal approach in all applications and should be used with care.

*Reliability as a Function of Number of Items and MIC*

To illustrate effects of the number of items and the MIC on scale reliability, we calculated the reliability for various combinations of these factors for hypothetical scales using the Spearman–Brown prophecy formula (McDonald, 1999). As shown in Table 3.2, the reliabilities of scales of different length are shown for original scales with 10, 20, 30, or 40 items, crossed by levels of MIC that span the approximate range discussed by Clark and Watson (1995)—specifically, MIC levels of .2, .3, .4, and .5. When researchers use a short form, they typically look first at the number of items and the homogeneity coefficient or reliability of the long form; these are shown in the first and third columns of Table 3.2, respectively, along with the MIC among items consistent with the associated level of reliability. Each row in Table 3.2 has three additional

TABLE 3.2
Predicted Reliabilities for Short Forms as a Function of Properties of Long Form

| Long form | | | Length of short form as function of long form | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 75% as long | | 50% as long | | 25% as long | |
| No. of items | MIC | $r_{xx}$ | No. of items | $r_{xx}$ | No. of items | $r_{xx}$ | No. of items | $r_{xx}$ |
| 40 | .50 | .98 | 30 | .97 | 20 | .95 | 10 | .91 |
| | .40 | .96 | | .95 | | .93 | | .87 |
| | .30 | .94 | | .93 | | .90 | | .81 |
| | .20 | .91 | | .88 | | .83 | | .71 |
| 30 | .50 | .97 | 23 | .96 | 15 | .94 | 8 | .89 |
| | .40 | .95 | | .94 | | .91 | | .84 |
| | .30 | .93 | | .91 | | .87 | | .77 |
| | .20 | .88 | | .85 | | .79 | | .67 |
| 20 | .50 | .95 | 15 | .94 | 10 | .91 | 5 | .83 |
| | .40 | .93 | | .91 | | .87 | | .77 |
| | .30 | .90 | | .87 | | .81 | | .68 |
| | .20 | .83 | | .79 | | .71 | | .56 |
| 10 | .50 | .91 | 8 | .89 | 5 | .83 | 3 | .75 |
| | .40 | .87 | | .84 | | .77 | | .67 |
| | .30 | .81 | | .77 | | .68 | | .56 |
| | .20 | .71 | | .67 | | .56 | | .43 |

*Note.* MIC = mean interitem correlation. Tabled values are theoretical estimates derived from MIC and number of items using the Spearman–Brown prophesy formula. A Microsoft Excel spreadsheet with other combinations of conditions is available at http://www.Quant.KU.edu/resources/published.html.

pairs of columns, one pair of columns for a scale 75% as long as the original form (i.e., discarding one fourth of the items), a second pair for a scale 50% as long (i.e., discarding half of the items), and a final pair for a scale 25% as long (i.e., discarding three fourths of the items). Thus, a 40-item scale with an MIC of .3 would have a reliability of .94, and a 20-item scale with MIC of .2 would have a reliability of .83.

Values in Table 3.2 appear to present a rather positive picture, with relatively high levels of reliability in many parts of the table. But one should remember that most original (or long-form) scales used in psychology have between 10 and 20 items per dimension. Suppose a researcher wanted to keep reliability above .80 and knew that the reliability of a 20-item scale was .90. This level of reliability would arise from an MIC of around .3, and the researcher would expect a short form containing half of the items from this scale (10 items) to have a reliability of .81. If this level of reliability were deemed too low, then keeping more items would be advisable. Or, if the original form of a 10-item scale had a reliability of .87 (i.e., MIC of .4), deleting more than about three items from the scale would lead to reliability below .80, which may be unacceptable.

## FUNDAMENTAL ISSUES IN PSYCHOLOGICAL MEASUREMENT: VALIDITY

Reliability concerns precision of measurement (i.e., how well the set of items measures whatever it is they assess). *Validity*, in contrast, concerns whether the scale assesses the construct it was designed to measure. Traditionally, methodologists have discussed a tripartite conception of validity—content validity, criterion validity, and construct validity—although more recent work (e.g., Messick, 1995) has extended these notions in interesting ways.

*Content validity* refers to how well a scale embodies content from all definable aspects of the domain to be assessed. Thus, a math achievement test for second graders may contain simple and complex addition items, simple and complex subtraction items, and word problems that can be translated into simple numerical operations, but it should not contain multiplication or division items or addition and subtraction items with decimals and fractions because these operations are not usually covered in the second grade curriculum. Content validity is typically assessed by asking subject-matter experts whether the measure is sufficiently comprehensive of all aspects of a domain. When using secondary data, short forms will often have been constructed from longer, original measures, and the researcher should ensure that the same breadth of content is shown in the short form as in the original form. If breadth of coverage is limited, then the short form may measure a somewhat narrower construct than assessed by the full instrument, and this should be noted.

*Criterion validity* is established by examining the predictive or concurrent correlations of a focal scale with key variables that are identified as criteria. Grades in school and scores on established intelligence tests are often used as criteria for new intelligence tests; job performance ratings might be used as criteria when evaluating the criterion validity of a battery of measures used in personnel selection. The magnitude and direction of these criterion-related correlations should be consistent with theory and past research with similar constructs. The stronger the correlation of a scale with a criterion measure, the stronger the criterion validity of the scale. For example, an intelligence test is likely to have good criterion validity for predicting educational attainment, because educational attainment is a near measure for an intelligence test, where *near* refers to a close connection theoretically and, perhaps, temporally. In contrast, an intelligence test may have rather lower levels of validity for predicting a criterion such as annual salary, which would be a far measure for the intelligence test (i.e., where *far* indicates much less close connection theoretically and, probably, temporally).

Finally, *construct validity* concerns whether a scale is a good measure of the theoretical construct it was designed to measure. No single study can establish the construct validity of a scale; instead, construct validity is gauged

by the pattern of results obtained across all studies using the scale. This pattern should satisfy several criteria: (a) the scale should correlate highly with other, well established measures of the same construct; (b) the scale should correlate much lower with measures of quite different constructs; and (c) scale scores should vary as a function of relevant contexts or conditions. To reiterate, the correlation of a scale with measures of other constructs need not always involve strong positive correlations, but rather can involve a well reasoned set of hurdles that include zero, small, medium, and large correlations in both positive and negative directions—all of which should be consistent with theory underlying the measure and its expected levels of correlation with other measures. In their description of a hypothetical measure of social intelligence, Westen and Rosenthal (2003) argued that it should correlate moderately positively with verbal IQ ($r = .50$), at a low positive level with Extraversion ($r = .10$), and moderately negatively with hostile attribution bias ($r = -.40$). Even using the elegant Westen and Rosenthal (2003) approach, construct validity of a scale cannot be captured with a single set of correlations but is summarized by examining all evidence that has accrued using the scale. A scale may have good construct validity for certain inferences but much poorer validity for others. Thus, construct validity is not an all-or-nothing affair but requires an understanding and summary of available research on the scale and considerations of the uses to which the scale will be put.

Much psychological research suffers an unfortunate confirmation bias (Widaman, 2008), whereby researchers hypothesize that relatively high, significant correlations will hold between certain measures and pay less attention to variables with which the focal measure should correlate at low levels. When correlations are evaluated, correlations hypothesized to differ significantly from zero are treated as theoretically important even if they are relatively small in magnitude, and correlations hypothesized to be negligible are treated as essentially equal to zero, even if they barely miss being deemed statistically significant. To combat this unfortunate bias, researchers should develop hypotheses regarding both the convergent and discriminant validity of their measures, as Westen and Rosenthal (2003) argued. *Convergent validity* refers to the degree to which a set of measures converges on the construct of interest. Convergent validity is supported if measures of the same purported construct exhibit high intercorrelations. *Discriminant validity* describes the degree of meaningful separation, or lack of substantial correlation, between indicators of putatively distinct constructs. Discriminant validity is supported if relations between different constructs are approximately zero in magnitude or at least much smaller than convergent correlations for the measures. The classic article by Campbell and Fiske (1959) should be consulted regarding convergent and discriminant validation, as should more recent work on structural equation modeling (SEM) of such data (e.g., Eid & Diener, 2006; Widaman, 1985).

## Relations Between Reliability and Validity

A common axiom of psychometric theory is that a scale cannot be valid if it is not reliable. If reliability is the proportion of variance in a measure that is true score variance, and validity is an index of how highly a measure correlates with the construct it purports to measure, then the square root of reliability places an upper limit on the validity of the measure. A scale with no reliable variance should be unable to correlate with any other measure. Because reliability places an upper bound on validity, rules of thumb have been offered for acceptable levels of reliability for measures. By convention, researchers often state that reliabilities between .70 and .80 are acceptable for research purposes, between .80 and .90 are preferred for research purposes, and above .90 (and preferably above .95) are required for individual assessment and diagnostic purposes. A scale with a reliability of .70 would still have a maximal validity of .84 if the criterion were perfectly reliable (McDonald, 1999, p. 133).

Classical tools for investigating statistical relations between variables, including regression analysis and analysis of variance, require the assumption that all variables are measured without error. All associated estimation theory—including estimating the strength of relations between variables, parameter estimates, standard errors, and confidence intervals of estimates—is based on the crucial assumption of perfect reliability. If measures have less than perfect reliability, parameter estimates are generally reduced in magnitude, standard errors and associated confidence intervals will be increased, and these effects will influence Type I and Type II error rates. Specifically, *Type I error rates*—or the likelihood that one would conclude that a relation between variables is significant when in fact it is due entirely to chance—would be lower as a function of lowered reliability of measures, so tests of significance would be negatively biased. However, *Type II error rates*—or the likelihood that one would fail to find significance for a relation when in fact it differed from zero—would increase as a function of lowered reliability of measures. Clearly, higher reliabilities are better, and the closer the reliabilities of measures are to unity, the more closely actual Type I and Type II error rates approximate their nominal values.

# RECOMMENDATIONS: HOW TO CREATE AND EVALUATE A SHORT FORM OPTIMALLY

When creating a short form of a scale, the ultimate issue should be the validity of the short form, rather than its reliability (John & Soto, 2007). That said, reliability should not be disregarded; indeed, because reliability is a prerequisite for validity, reliability should be the first psychometric index to

be evaluated. However, any short form will have fewer, often many fewer, items than the original scale, so reliability of a short form is likely to be appreciably lower than that for the full scale.

If a researcher were interested in creating a short form that has as high a level of reliability as possible, he or she might select the subset of items that have the highest MIC, because higher reliability arises from higher MIC values. But if selection of only items with high MIC leads to a biased selection of items (i.e., failure to preserve the breadth of the domain across the items in the short form), then the validity of the short form may be severely compromised, even as the reliability of the short form is maximized. Stated differently, the optimal set of indicators for a short form measure of a given construct may not be the indicators that have the highest internal consistency from among the possible items of the full scale; in fact, maximizing internal consistency can lead to suboptimal outcomes. Indeed, Loevinger (1954) described the attenuation paradox in which increasing reliability leads to increasing validity up to a point, beyond which point further increases in homogeneity reliability decreases validity. Or, selecting items that correlate most highly can lead to selection of items with extreme levels of item content overlap, leading to bloated specific factors that represent pairs of redundant items (Cattell & Tsujioka, 1964). Researchers must take care when developing short forms from longer original measures because common approaches for developing short forms are potentially problematic.

The most common methods for constructing short form measures are (a) selecting a subset of items with the highest MIC (described earlier), to maximize reliability of the short form; (b) selecting items with the highest loadings on the common factor underlying the items, to obtain items most closely aligned with the factor; (c) selecting items with the highest correlation with the total scale score (preferably the highest correlation with a composite of the remaining items on the scale); (d) selecting items with the highest face validity, or items that are the most obvious indicators of the construct; or (e) selecting items randomly from the original scale. Each of the preceding methods has flaws, and most methods have several. Methods (a), (b), and (c) use empirical methods, basing decisions on patterns of results from a particular set of data. Because the subset of items that appears to be optimal might vary across different sets of empirical data, basing item selection on a single data set is problematic and capitalizes on chance results in a single sample. Further, Methods (a) through (d) may result in a narrowing of item content, restricting improperly the breadth of the item content in the full scale. Method (d) is based on subjective judgments by researchers, and care must be taken lest the predilections of one researcher bias the item selection in idiosyncratic ways. Finally, Method (e) appears to be an unbiased approach to item selection, but researchers usually want to select the best items for a short

form, not a random sample of items. Only if all items of the larger scale were equally good would a random selection of a subset of items be a reasonable approach. In practice, items are rarely equally good, so this approach probably would not lead to an optimal short form.

An underused approach that has potential merit is to identify a subset of items that maintains the factorial integrity of the construct. By *factorial integrity*, we mean that the construct maintains its levels of association with a select set of criteria or other constructs and that the estimated mean and variance of the construct are minimally changed. This focus on factorial integrity of the construct is a focus on validity—ensuring that the construct embodied in the short form maintains the same position in the nomological network of relations among constructs (cf. Cronbach & Meehl, 1955) as did the full, longer scale. SEM can be used iteratively to identify the subset of items that maintains factorial integrity of the short form. Here, one fits a model using the items from the full scale to represent the construct and includes a carefully chosen set of additional criteria. In a second analysis, all aspects of the model in the first analysis are identical except that one selects a subset of items to represent the focal construct. The mean, variance, and associations of the construct based on the full scale are compared with the mean, variance, and associations of the construct based on the selected subset of items. This model would be iteratively fit until an optimal subset of items is identified that maintains the factorial integrity of the construct.

When using secondary data, using the preceding methods to construct short form measures of constructs may not be possible. Secondary data are what they are—existing data that can be used for new purposes. As a result, if short forms were used when the data were collected, then existing short forms have already been constructed, and the user must live with those existing short forms. However, many secondary data sets have large selections of items that were never assigned to a priori scales. Instead, the questions probe various domains of content, and individual questions may have been used in prior research to answer particular questions. Nothing should stop the enterprising researcher from using these items to create scales to represent constructs of interest, but care must be taken when doing so. Of course, creating new scales from collections of items in an existing data set will involve new scales, not short forms of established scales. Still, the resulting new scales will likely consist of a fairly small number of items, so all principles and concerns related to analysis and evaluation of short forms still apply.

Additional steps can be pursued to explore the use of short forms constructed through the preceding steps. One step would be to perform a factor analysis to determine whether factors aligned with newly constructed short forms can be confirmed in the secondary data. Researchers should ensure that common factor techniques are used, because the use of principal-components

analysis—the chief alternative to common factor analysis and the default in most computer programs—can lead to substantial bias in loadings and other parameter estimates, especially when the number of items analyzed is not large (Widaman, 1993, 2007). Factor analyses can be conducted at the item level, but item-based analyses are often problematic because of the issue of bloated specifics noted above. One useful alternative is the use of item parcels, which are sums of subsets of items composing a scale (Kishton & Widaman, 1994; Little, Cunningham, Shahar, & Widaman, 2002). Suppose one sifted through a set of items in an existing data set with the intent of developing short scales for Big Five dimensions of personality. Suppose as well that one identified a total of 30 items, six items for each of the five dimensions, that seemed to capture the respective dimensions reasonably well. Item-based analyses could be pursued, but a researcher would also be justified in forming 3 two-item parcels for each scale and then performing the factor analysis on the set of 15 two-item parcels. For additional suggestions on use of factor analysis in revising and evaluating measures, see Floyd and Widaman (1995) and Reise, Waller, and Comrey (2000).

## RECOMMENDATIONS: WHAT TO DO WHEN WORKING WITH ALREADY-CREATED SHORT FORMS

When using existing data, researchers hope to find measures of key constructs needed to answer theoretical questions, even if some measures are short forms. If short forms of existing instruments are present, the user must evaluate the psychometric properties of the short forms and then analyze data with these properties in mind. We have several recommendations for data analyses that are informed by the properties of the scales analyzed.

The first recommendation is to estimate the reliability of each scale from the existing data set that will be used in the current research. Do not assume that original levels of reliability will be obtained, particularly if the scale is a short form composed of many fewer items than in the original form. Reliability is usually estimated using a homogeneity coefficient, and we recommend coefficient omega over the more commonly used coefficient alpha because it relies on more reasonable assumptions regarding the items on a short form—that items have a congeneric structure, rather than tau equivalent structure. Also, if coefficients omega and alpha diverge in magnitude, coefficient alpha is likely a biased underestimate of scale reliability, leading to biased overcorrection for unreliability when using the correction for attenuation. If data have a longitudinal component, then the correlation of a given scale from one measurement occasion with the same scale at the next measurement occasion can be used to estimate test–retest reliability.

As a second recommendation, when investigating correlations among measures, investigators should use the correction for attenuation to correct all correlations for unreliability and then evaluate both the raw and disattenuated correlations. The disattenuated correlation between variables X and Y provides an estimate of the correlation between the true scores of X and Y, and is calculated as

$$r_{XYc} = \frac{r_{XY}}{\sqrt{r_{XX}\, r_{YY}}}, \tag{3}$$

where $r_{XYc}$ is the correlation between X and Y corrected for attenuation due to unreliability, $r_{XY}$ is the Pearson product–moment correlation between the observed scores on variables X and Y, and $r_{XX}$ and $r_{YY}$ are the reliability coefficients for variables X and Y, respectively. Because any short form is likely to have somewhat lower reliability than the full form from which it was derived, the lowered reliability of the short form will lead to lower correlations of the short form with other variables. Correcting correlations for attenuation due to unreliability will allow one to evaluate the influence of lowered reliability on estimated relations among variables.

Our third recommendation is to consider analyzing data using latent-variable SEM, which automatically corrects for attenuation due to unreliability and also corrects for attenuation due to specific variance. The correction for attenuation obtained using SEM approaches arises because latent variables in such models represent error-free constructs, and relations among latent variables are relations from which measurement error and specific variance have been partialed out (Little, Lindenberger, & Nesselroade, 1999). When latent-variable SEM is used, whether one used a short form or the long form of a scale should not matter, at least in theory. Estimated relations among latent variables should be comparable across long and short forms because the analysis accommodates the different levels of reliability of the short and long forms. Of course, this contention is based on several assumptions, a chief one of which is that the short form provides an unbiased, if less reliable, measure of the construct assessed by the long form. If item selection in deriving the short form from the long form of the scale resulted in any narrowing or bias in item content of the short form, then the equality of relations for short and long forms need not hold (for researchers with little acquaintance with SEM, see Kline, 2004).

Our fourth recommendation is that to the degree possible, a researcher should investigate the validity of a construct by inspecting criterion-related associations of the construct with other measures in the data set. Here, one would examine other published data that used the full scale and note the associations of the full scale with as many other constructs as possible. If

the existing data set has similar constructs or criteria, the short form should show patterns of association with these variables that are sufficiently similar to encourage further consideration. Admittedly, differences between studies can result in changes in correlations among variables, but similar patterns of correlations among constructs should tend to hold across studies.

Our fifth recommendation, which pertains specifically to longitudinal studies, is that researchers take care to ensure that variables are on the same metric across times of measurement. In longitudinal data sets, items in a short form may change from one occasion to the next. In such situations, researchers must ensure that measurements are on the same underlying scale if growth or change is the object of study. Simple approaches—such as the computation of average item scores or of proportion scores—rest on problematic assumptions that the items from the different forms function in precisely the same fashion in assessing the construct. Thus, these simple approaches are too simpleminded and problematic for current scientific work.

Linking the metric of latent variables across time in the presence of changes in the sets of items on a short form can be accomplished using either SEM or item response theory (IRT) approaches. Many different scenarios for the migration of items off of or onto short forms can be envisioned. For example, under one scenario, Items 1 through 12 (a short form of a 30-item instrument) might be used to assess a construct for three times of measurement during early adolescence; as participants move into later adolescence and are assessed three additional times, the first six items are dropped and Items 13 through 18 are substituted for them. Thus, 12 items are used at each measurement occasion, and one subset of six items (Items 7–12) is used at all measurement occasions. Under a second scenario, Items 1 through 12 are included at the first three times of measurement, Items 1 through 12 are supplemented with Items 13 through 24 at the fourth time of measurement, and all remaining times of measurement involve only Items 13 through 24. Under this scenario, no core set of items is administered across all measurement occasions, but all items that appear at any time of measurement are used at the fourth time of measurement. Clearly, many additional scenarios could be posed as likely to occur in existing data.

Under the first scenario, an SEM approach might use Items 7 through 12 to define two parcels, the same items would be assigned to these two parcels at each time of measurement, and these two parcels would appear at all times of measurement. Items 1 through 6 could be summed to form a third parcel for measurement occasions 1 through 3, and Items 13 through 18 could be summed for a third parcel at the last three measurement occasions. If the factor loadings, intercepts, and unique variances for the two common parcels were constrained to invariance across all times of measurement, the resulting latent variables would be on a comparable scale across all times of measurement. An appropriate IRT approach would have a similar rationale, requiring the

presence of all 18 items in one analysis, and resulting theta scores (which are estimates of participant level on the construct) would be on a comparable scale across time.

The second scenario is, in some ways, simpler than the first, with all items that are used at any time of measurement appearing at a single occasion of measurement (the fourth time of measurement). The key analyses would be the linking of scores across the two forms—Items 1 through 12 and Items 13 through 24—in analyses using data from the fourth measurement occasion. Then, whether using SEM or IRT approaches, invoking invariance of parameter estimates from the fourth occasion of measurement on corresponding estimates at other occasions of measurement would lead to latent-variable scores on the same metric. Details of these methods are beyond the scope of the present chapter. The Embretson and Reise (2000) text offers a very good introduction to IRT procedures in general, and recent work (e.g., Cho, Boeninger, Masyn, Conger, & Widaman, 2010; Curran et al., 2008) provides relevant details and comparisons between SEM and IRT approaches.

## CONCLUSIONS

Creating and using short forms of longer measurement instruments is a fact of life in many areas of psychology. Sometimes short forms are used to yield optimal screening instruments; other times they are used to incorporate measures of many constructs in a single protocol that is not too long. The key question to address with short forms is this: Has the reduced form of the scale undermined its validity for your intended purpose (John & Soto, 2007)? Our discussion has covered multiple ways in which a short form can compromise validity. The most obvious reason is the decrease in reliability that should occur with the decrease in the number of items, an effect embodied in the Spearman–Brown prophecy formula that was proposed 100 years ago (Spearman, 1910; Brown, 1910). But the validity of a short form can also be diminished by biased selection of items from the longer form, thereby narrowing the range or breadth of content covered by the scale and changing fundamentally the nature of the underlying dimension tapped by the scale (see Little et al., 1999). Thus, short forms are not a panacea for research in psychology, and researchers should be careful when selecting short forms of longer scales for their own research.

However, secondary data are what they are and cannot be changed. They often also have notable advantages, such as large sample sizes and probability-based sampling plans. If the data contain short forms, then these short-form scales should be handled with the most optimal mathematical and statistical techniques available. Many existing data sets are invaluable; they can be used

to answer questions of current theoretical interest, and new data with long forms of scales would take years or decades to gather anew. Thus, rather than ruing the absence of long form instruments, we recommend that researchers concentrate instead on the most appropriate and state-of-the-art ways to analyze the existing data, warts and all.

Our strongest recommendations range from tried-and-true to innovative approaches to data analysis that can and should be used with short-form instruments. The tried-and-true methods include estimation of reliability in the sample of data at hand and use of the correction for attenuation when estimating relations among variables. Similarly, SEM methods, particularly multiple-indicator SEMs, accomplish a great deal in terms of correcting for poorer measurement properties of short forms (Little et al., 1999), and these methods are not generally novel any more. However, the ways in which SEM or IRT can be used to ensure that the scale of a latent variable remain the same across measurement occasions in the face of changes in the composition of a short form are innovative. Current research is being done to illustrate how this can and should be done and to establish optimal procedures for meeting these analytic goals.

We have considered likely outcomes when short forms are used, offered basic ideas about how to gauge the psychometric properties of short form data, provided some guidelines about constructing new scales from older collections of items in existing data, and recommended analytic strategies for evaluating short form data and including them in models. Valuable secondary data are out there, often containing short form instruments but waiting to be used as the unique basis for answering interesting, crucial, state-of-the-science questions. We encourage researchers to exploit such resources, using analytic approaches that are appropriate for the data and provide optimal tests of their conjectures.

## FOR FURTHER READING

Informative articles on reliability, validity, and scale development and revision that we recommend include

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319. doi:10.1037/1040-3590.7.3.309

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286–299. doi:10.1037/1040-3590.7.3.286

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287–297. doi:10.1037/1040-3590.12.3.287

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353. doi:10.1037/1040-3590.8.4.350

Informative introductions to psychometrics, structural modeling, and IRT that we recommend include

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Kline, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

McDonald, R. P. (1999). *Test theory*. Mahwah, NJ: Erlbaum.

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

## REFERENCES

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *The British Journal of Psychology, 3*, 296–322.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0046016

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement, 24*, 3–30. doi:10.1177/001316446402400101

Cho, Y. I., Boeninger, D. K., Masyn, K. E., Conger, R. D., & Widaman, K. F. (2010). *Linking of scales in longitudinal research: Comparing item response theory and second-order latent growth model approaches*. Manuscript submitted for publication.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319. doi:10.1037/1040-3590.7.3.309

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstances for POMP. *Multivariate Behavioral Research, 34*, 315–346. doi:10.1207/S15327906MBR3403_2

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi:10.1037/h0040957

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44*, 365–380. doi:10.1037/0012-1649.44.2.365

Eid, M., & Diener, E. (Eds.). (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association. doi:10.1037/11383-000

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286–299. doi:10.1037/1040-3590.7.3.286

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big Five Inventory—Versions 4a and 5. Berkeley, CA: Institute of Personality and Social Research.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461–494). New York, NY: Guilford Press.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757–765. doi:10.1177/0013164494054003022

Kline, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173. doi:10.1207/S15328007SEM0902_1

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*, 192–211. doi:10.1037/1082-989X.4.2.192

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493–504. doi:10.1037/h0058543

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*, 1–21.

McDonald, R. P. (1999). *Test theory*. Mahwah, NJ: Erlbaum.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (4th ed.). Los Angeles, CA: Authors.

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

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*, 203–212. doi:10.1016/j.jrp.2006.02.001

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287–297. doi:10.1037/1040-3590.12.3.287

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353. doi:10.1037/1040-3590.8.4.350

Spearman, C. (1910). Correlation calculated with faulty data. *The British Journal of Psychology, 3*, 271–295.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84*, 608–618. doi:10.1037/0022-3514.84.3.608

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9*, 1–26. doi:10.1177/014662168500900101

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28*, 263–311. doi:10.1207/s15327906mbr2803_1

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177–203). Mahwah, NJ: Erlbaum.

Widaman, K. F. (2008). Integrative perspectives on cognitive aging: Measurement and modeling with mixtures of psychological and biological variables. In S. M. Hofer & D. F. Alwin (Eds.), *The handbook of cognitive aging: Interdisciplinary perspectives* (pp. 50–68). Thousand Oaks, CA: Sage.

Zinbarg, R. E., Revelle, W., & Yovel, I. (2007). Estimating $\omega_h$ for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement, 31*, 135–157. doi:10.1177/0146621606291558

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating the generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators of $\omega_h$. *Applied Psychological Measurement, 30*, 121–144. doi:10.1177/0146621605278814

# SECONDARY DATA ANALYSIS

## An Introduction for Psychologists

Edited by
Kali H. Trzesniewski, M. Brent Donnellan,
and Richard E. Lucas

American Psychological Association • Washington, DC