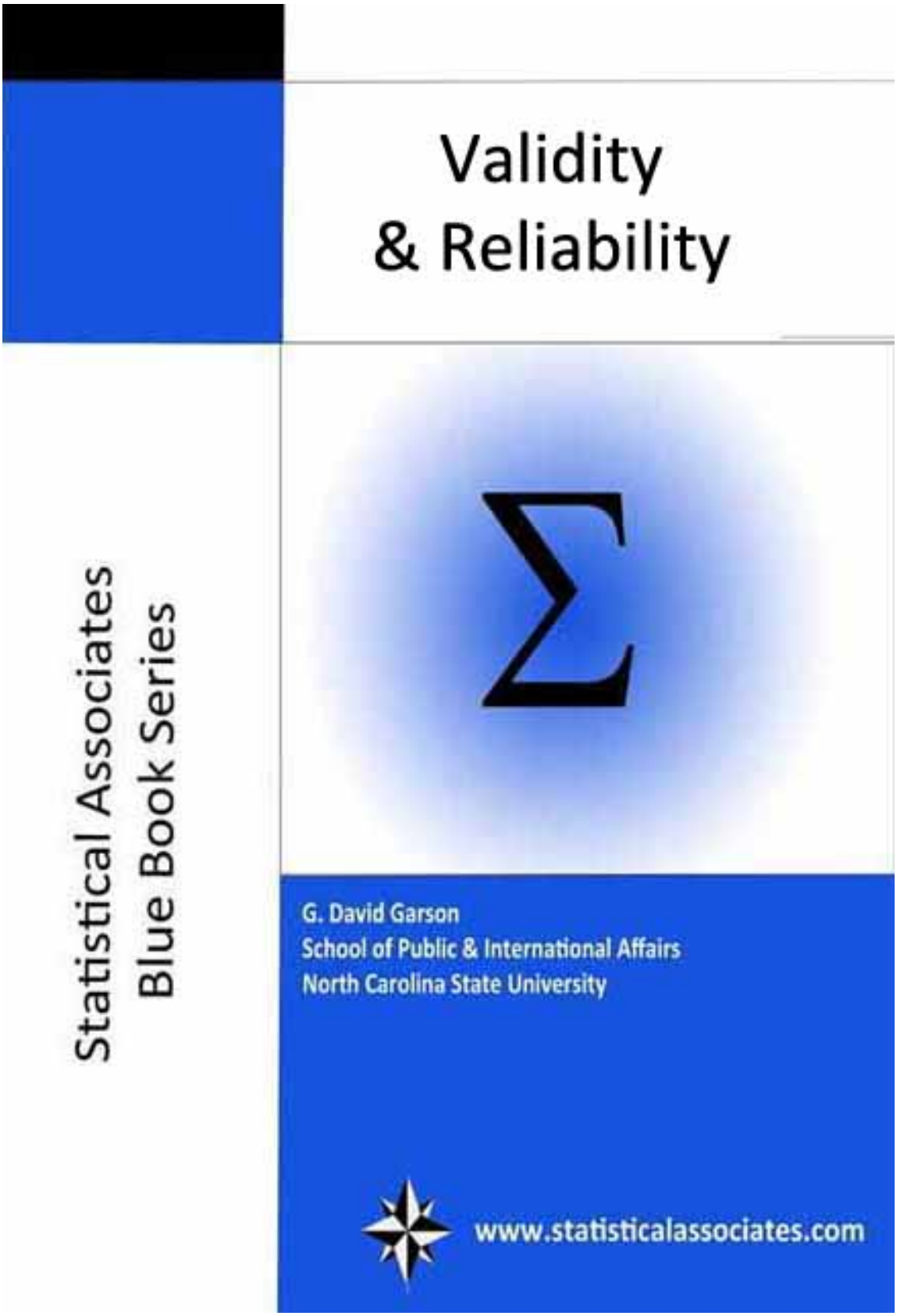


ISBN: 978-1-62638-005-9

Contact:

G. David Garson, President
Statistical Publishing Associates
274 Glenn Drive
Asheboro, NC 27205 USA

Email: gdavidgarson@gmail.com
Web: www.statisticalassociates.com

Table of Contents

Validity and Reliability

VALIDITY OVERVIEW

A study is valid if its measures actually measure what they claim to, and if there are no logical errors in drawing conclusions from the data. There are a great many labels for different types of validity, but they all have to do with threats and biases which would undermine the meaningfulness of research. Researchers disagree on the definitions and types, which overlap. The typology is much less important than understanding the types of questions the researcher should ask about the validity of research.

The question of validity arises in three contexts.

1. At the level of the item or measure, the researcher is concerned with the content itself (face validity), unidimensionality (ex., avoiding double-headed items like the agree/disagree item, "The head of my organization is inspiring and effective."), and whether the item correlates as expected (criterion validity)
2. At the level of the construct, the researcher is concerned with whether the indicator measures cohere well (convergent validity) and are differentiated from indicator measures for other constructs in the model (divergent validity). Also of concern is whether the construct correlates as expected (criterion validity) and does not reflect an instrumentation artifact (common method bias).
3. At the level of the study, the researcher is concerned with whether statistical procedures have been applied properly (statistical validity) and whether research design is sound (internal validity) and whether generalizations are appropriate (external validity).

Reliability, discussed further below, is the correlation of an item, scale, or instrument with a hypothetical one which truly measures what it is supposed to. As the "true measure" is not available, reliability must be estimated by correlation with what is assumed to be true. All reliability coefficients are forms of correlation coefficients, but there are multiple types representing different meanings of

reliability. One type, for instance, is "internal consistency reliability" which assumes that if all items in a scale truly measure the same thing, they should be highly intercorrelated with each other. Another is "split-half reliability," which assumes that if all items in a scale truly measure the same thing, then a randomly selected set of half the items should correlate highly with another randomly selected set.

All valid measures are reliable but not all reliable items are valid. For instance, the split-half reliability method may show statistical reliability yet experts in the field of study may feel the items do not measure what they are supposed to ("face validity" is lacking). Other forms of correlation are often used to establish that a reliable measure is also valid. One such validity coefficient, for instance, is the correlation of a measure with another measure which is well-established and accepted in the field as a measure of the same or a similar thing (this is called "criterion validity").

Validity: Historical background

Some early writers simply equated validity with establishing that a construct's scale correlated with a dependent variable in the intended manner. In this early view, a scale might be considered valid as a measure of anything with which it correlated (Guilford 1946).

Types of validity were codified in 1954 by the American Psychological Association, which identified four categories: content validity, construct validity, concurrent validity, and predictive validity (APA, 1954). Each type corresponded to a different research purpose:

1. Content validity had to do with subject-matter content testing
2. Construct validity concerned measuring abstract concepts like "confidence"
3. Concurrent validity dealt with devising new scales or tests to replace existing ones
4. Predictive validity focused on devising indicators of future performance

A 1966 update to the APA typology combined the last two types under the label criterion-related validity (APA, 1966).

Later, Sheperd (1993) was among those who argued that both criterion and content validity were subtypes of construct validity, leaving only one type of validity. This unified view of validity supported the notion that only rarely could a researcher establish validity with reference to a single earlier type. Moreover, Cronbach's (1971: 447) earlier argument that validity could not be established for a test or scale, only for interpretations researchers might make from a test or scale, also became widely accepted in the current era. Some, such as Messick (1989), accept construct validity as the only type, but argue for multiple standards for assessing it:

- Relevant content, based on sound theory or rationale
- Internally consistent items
- External correlation with related measures
- Generalizability across populations and time
- Explicit in its social consequences (ex., racial bias).

Construct validity is sometimes also called factorial validity. It has to do with the logic of items which comprise measures of concepts (constructs). A good construct has a theoretical basis which is translated through clear operational definitions involving measurable indicators. A poor construct may be characterized by lack of theoretical agreement on its content, or by flawed operationalization such that its indicators may be construed as measuring one thing by one researcher and another thing by another researcher.

A construct is a way of defining something, and to the extent that a researcher's proposed construct is at odds with the existing literature on hypothesized relationships using other measures, its construct validity is suspect. For this reason, the more a construct is used by researchers in settings with outcomes consistent with theory, the greater its construct validity. For their constructs, researchers should at a minimum establish both of the two main types of construct validity, convergent and discriminant, discussed below.

In a nutshell, over the last half century the concept of validation has evolved from establishing correlation of a measure with a related dependent variable to the idea that researchers must validate each interpretation of each scale, test, or instrument measuring a construct and do so in multiple ways which only taken together form the whole of what validity is.

The remainder of this work largely accepts the unified view of validity, centering on construct validity, it organizes discussion around types of validity coefficients and validation procedures used by researchers.

Convergent validity

Do items in a scale converge on a unidimensional meaning?

Often a researcher will have multiple indicator items for each construct in his or her model -- for example, 10 survey items measuring "confidence". This is standard practice because a construct based on multiple indicators is thought to be more stable and representative than a single-item measure of the same construct. However, it is possible the 10 items actually measure different constructs, not all measuring the same construct. The researcher mush show that the items in the scale converge, indicating that a single dimension of meaning is being measured.

Convergent validity is assessed by the correlation among items which make up the scale or instrument measuring a construct (internal consistency validity), by the correlation of the given scale with measures of the same construct using scales and instruments proposed by other researchers and already accepted in the field (criterion validity). Convergent validity is also assessed by correlation of across samples (ex., a racial tolerance scale using subject data should correlate with the scale using spousal data) or across methods (ex., survey data and archival data). One expects these correlations to be at least moderate to demonstrate convergent validity.

Cronbach's alpha as a validity coefficient

Internal consistency is a type of convergent validity which seeks to assure there is at least moderate correlation among the indicators for a concept. Poor convergent validity among the indicators for a construct may mean the model needs to have more factors.

Cronbach's alpha is commonly used to establish internal consistency construct validity, with .60 considered acceptable for exploratory purposes, .70 considered adequate for confirmatory purposes, and .80 considered good for confirmatory

purposes. Cronbach's alpha is both a validity coefficient and a reliability coefficient, discussed further below in the section on reliability.

Example: In their study of direct-to-consumer (DTC) advertising, Huh, Delorme, and Reid (2006) developed consumer attitude constructs, which they validated by reporting Cronbach's alpha levels of 0.87, 0.88, and 0.89, respectively. See Huh, J, Delorme, D. E., & Reid, L. N. (2006). Perceived third-person effects and consumer attitudes on prevetting and banning DTC advertising. *Journal of Consumer Affairs* 40(1): 90.

Other convergent validity criteria

Simple factor structure

Simple factor structure is another test of internal consistency, seeking to demonstrate for a valid scale that indicator items for a given construct load unambiguously on their own factor. This tests both convergent and discriminant validity, as discussed in the separate Statistical Associates "Blue Book" volume on "Factor Analysis". The usual rule-of-thumb criterion is that simple factor structure exists to the extent that the proposed scale items all load most heavily on the same factor, and that they do not cross-load heavily on other factors. The common cut-offs are that intended scale items should load at the .70 level or higher on their factor, and that all cross-loadings should be below .30. As the simple factor structure criterion is stringent, some researchers will accept that convergent validity has been established as long as 95% of loadings conform to simple factor structure cut-offs.

In the usual version of this approach, all items for all scales are factored together. In a more stringent version, indicator items for each pair of constructs are factored separately to determine if they all load on a single general factor for that construct.

Rasch models

In brief, when Cronbach's alpha or simple factor structure are used to validate the inclusion of a set of indicator variables in the scale for a construct, the researcher is assuming a linear, additive model for ordinal indicator items. Linearity is assumed as part of correlation, which is the basis for clustering indicator variables

into factors. Additivity is also assumed, meaning that items will be judged to be internally consistent only if they are mutually highly correlated. However, items may lack high intercorrelation but have a strong ordered relationship (ex., a scale of math ability composed of items of ascending difficulty). For this reason, many researchers prefer to use a Rasch model to guide scale construction, in preference to additive models like Cronbach's alpha or factor analysis.

Rasch models, also called one-parameter logistic models, are an internal consistency test commonly used in item response theory for binary items, such as agree or disagreeing with a series of statements (though polytomous Rasch models are available also). Rasch models, like Guttman scales, establish that items measuring a construct form an ordered relationship (see Rasch, 1960; Wright, 1977, 1996). Note that a set of items may have ordered internal consistency even though they do not highly correlate (additive internal consistency, such as tested by Cronbach's alpha or factor structure). Ordered internal consistency reflects a difficulty factor, whereby answering a more difficult item predicts responses on less difficult items but not vice versa. The usual criteria for a good measure in Rasch modeling is a standardized infit standard deviation < 2.0 for both persons and items, demonstrating low misfit; separation > 1.0 for both persons and items, demonstrating sufficient spread of items and persons along a continuum; and reliability for both persons and items >= .7 for confirmatory purposes. However, Rasch analysis, which is beyond our scope here, is more complex than these cut-offs may suggest, enabling the researcher to analyze such things as whether a scale is good fit for certain groups of people but not for others.

On Rasch modeling in SPSS, see Tenvergert, Gillespie, & Kingma (1993). Ordered vs. ordinal scaling is discussed further in the separate Statistical Associates "Blue Book" volume on "Scales and Measures."

Average variance extracted (AVE)

Alternatively, and less commonly, the researcher may consider a construct to display internal consistency convergent validity if average variance extracted *(AVE)* is at least .50 (Chin, 1998; Höck & Ringle, 2006: 15). That is, variance explained by the construct should be greater than measurement error and greater than cross-loadings. Hair et al. (2010) state that AVE should be at least .50 and that the composite reliability (CR) should be greater than AVE. AVE is discussed further below in the section on discriminant validity. Composite

reliability also is discussed further below. For further discussion, see Fornell and Larcker (1981).

Common method variance

Common method variance is a type of spurious internal consistency which does not indicate convergent validity. Common method variance occurs when the apparent correlation among indicators or even constructs is due to their common source (spurious convergence). For instance, if the data source is self-reports, the correlation may be due the propensity of the subject to answer similarly to multiple items even when there is no true correlation among them. Common method variance is perhaps best assessed using structural equation modeling of error terms, discussed separately in the Statistical Associates "Blue Book" on "Structural Equation Modeling". Common method variance is sometimes also assessed by (1) factoring all indicators in the study to see if a single common factor emerges, indicative of common method variance; or (2) observing correlations between different indicators of the same construct using the same and different methods, with the expectation that these correlations will be high in the same method data and low in the cross-method data if common method variance is a problem. See Podsakoff, MacKenzie, Lee, & Podsakoff (2003).

Mono-method and/or mono-trait biases

Use of a single data-gathering method and/or a single indicator for a concept may result in bias. Various data-gathering methods have their associated biases (ex., the yea-saying bias in survey research, where people tell pollsters what they think they want to hear). In the same vein, it should be asked if the researcher has used randomization of items to eliminate order effects of the instrument, or established the unimportance of order effects? Likewise, basing a construct like "work satisfaction" on a single item dealing with, say, socializing with peers at work, biases the construct toward a particularistic meaning.

Multitrait-multimethod validation

In a multi-method, multi-trait validation strategy, the researcher not only uses multiple indicators per concept, but also gathers data for each indicator by multiple methods and/or from multiple sources. For instance, in assessing the concept of "tolerance," the researcher may have indicators for racial tolerance,

religious tolerance, and sexual orientation tolerance; and each may be gathered from the subject, the subject's spouse (assessing tolerance indicators for subject, not spouse), and the subject's parent (assessing tolerance indicators for subject, not parent). A correlation matrix is created in which both rows and columns reflect the set of three tolerance indicators, grouped in three sets -- once for subject data, once for spousal data, and once for parental data.

Discriminant validity

Do items in a two scales differentiate constructs?

Discriminant validity, also called divergent validity, is the second major type of construct validity. It refers to the principle that the indicators for different constructs should not be so highly correlated across constructs as to lead one to conclude that the constructs overlap. This would happen, for instance, if there is definitional overlap between constructs. Discriminant validity analysis refers to testing statistically whether two constructs differ (as opposed to testing convergent validity by measuring the internal consistency within one construct, such as by Cronbach's alpha, as discussed above).

Correlational methods

Correlational methods are considered to be among the less stringent tests of discriminant validity. A wide variety of correlational rules of thumb have been used in testing discriminant validity, however, a few of which are listed below.

- Reject an indicator if the item-scale correlation for its intended scale is less than its item-scale correlation with any other scale in the model.
- Reject an indicator if it correlates at an r = .85 level or higher with any indicator for any other scale in the model.
- Given two scales for two different constructs and given two criterion measures known to be associated with each construct, discriminant validity is upheld if the correlation of each scale with its own criterion measure is greater than its correlation with criterion measures for other scales in the model.
- Given two scales for two different constructs, low correlation between the scales (ex., < r=.3) demonstrates lack of definitional overlap and upholds discriminant validity.

Example. For a population of 211 demented and 94 mentally handicapped patients, Dijkstra, Buist, and Dassen concluded that the low correlations between the Scale for Social Functioning (SSF) full scale score and the other tested scales (the BOSIP Behavior Observation Scale for Intramural Psychogeriatrics) affirm the discriminant validity of the SSF scale. Dijkstra, A., Buist, G., & Dassen, T. (1998). A criterion-related validity study of the nursing care dependency. *International Journal of Nursing Studies* 35: 163-170.

Factor methods

Simple factor structure, discussed above, demonstrates discriminant as well as convergent validity. That is, discriminant validity is upheld if the factor loadings for each indicator load heavily (r >= .7) on the intended factor and cross-loadings are low (r < .3). Note factor structure will vary according to the factor model selected by the researcher. For further discussion, see Straub (1989).

Average variance extracted (AVE) method

An alternative factor-based procedure for assessing discriminant validity is that proposed by Fornell and Larcker (1981). In this method, the researcher upholds discriminant validity if AVE is greater than either maximum shared squared variance (MSV) or average shared squared variance (ASV) to demonstrate discriminant validity. For convergent validity, AVE should be .5 or greater and less than composite reliability (CR). The AVE for a factor or latent variable should also be higher than its squared correlation with any other factor or latent variable.

Note that as AVE and related validity coefficients are based on factor loadings, their values will vary according to the factor model. For further discussion, see Hair et al. (2010).

Computation of maximum shared squared variance (MSV) and average shared squared variance (ASV)

MSV is the square of the highest covariance for that factor while ASV is the average. See Hair et al. (2010) for further discussion.

Computation of AVE. AVE is the variance in indicator items captured by a construct as a proportion of captured plus error variance. AVE is calculated by the following steps if done manually.

1. In factor analysis, look at the standardized factor loadings table, which shows indicator items as rows and factors as columns. The factor loadings table comes from exploratory factor analysis (EFA), confirmatory factor analysis in structural equation modeling (CFA), or partial least squares modeling. Confirmatory factor analysis is the usual context. In AMOS and some other packages the factor loadings are labeled "standardized regression weights". All loadings of indicators intended to correspond to a given factor (in EFA) or latent variable (in CFA) should be strong (>=.7, though as low as .5 may be allowed for exploratory purposes) and significant.
2. Square the standardized factor loadings. The squared factor loadings are the communalities, also called the "item reliabilities". Sum these reliabilities for the items for any factor or latent variable. Let VE = variance extracted for a given factor = this sum for any given factor.
3. Average variance extracted = AVE = VE/number of items for the given factor. If 5 indicators are associated with a factor or latent variable, its AVE is its VE divided by 5. AVE should be at least .50. Anything less means that the variance explained is less than error variance.

Computation of construct reliability (CR)

CR, standing for construct or composite reliability, is calculated using the squared sum of standardized factor loadings and the sum of error variance. Error variance, called "delta" in SPSS and some other statistical packages, is 1.0 minus the item reliability (recall item reliability = the square of the standardized factor loading for an item). The numerator for CR is the sum of squared factor loadings as in step (2) above for AVE (that is, it is VE). The denominator is VE + the sum of error variances for the items in the factor. Let the error variance sum be EV. The formula for CR is VE/(VE + EV). CR will vary from 0 to 1.0, with 1.0 being perfect reliability. Its cut-offs are the same as for other measures of reliability: .6 is adequate for exploratory research, .7 is adequate for confirmatory research, and .8 or higher is good reliability for confirmatory research.

Example for AVE and CR

Below is a spreadsheet example of computing AVE and CR.

	A	B	C	D	E	F	G
1							
2							
3	FOR A GIVEN FACTOR OR LATENT VARIABLE						
4							
5	Indicator	Loading	Loading squared =		Delta=		t-value
6	Variable		Item reliabilities	AVE	1 - Reliability		
7							
8	Item1	0.538	0.289		0.711		23.752
9	Item2	0.412	0.170		0.830		23.372
10	Item3	0.414	0.171		0.829		23.493
11	Item4	0.49	0.240		0.760		27.877
12	Item5	0.254	0.065		0.935		14.07
13	Item6	0.217	0.047		0.953		11.529
14							
15	VE = SUM(C8:C13)		0.982				
16	AVE = VE/6			0.164			
17	EV = SUM(E8:E13)				5.018		
18	CR = VE/(VE+EV)					0.164	
19							
20	*All regression weights are standardized and significant.						
21							

Software

Modeling software such as AMOS, LISREL, and PLS all provide output for calculation of AVE, with free SmartPLS software doing so in a user-accessible way, illustrated below.

PLS

Quality Criteria

Overview

	AVE	Composite Reliability	R Square	Cronbachs Alpha
Incentives	0.827852	0.905812		0.792450
Motivation	0.823670	0.903310	0.430766	0.785929
SES	0.766229	0.867545		0.697499

	Communality	Redundancy
Incentives	0.827852	
Motivation	0.823670	0.305908
SES	0.766229	

In terms of presentation, it is customary to provide a matrix of squared covariances of each construct with each other construct, replacing the diagonal elements with the AVE for the column construct. If there is discriminant validity, then the diagonal element for a given column (construct) should be larger than any of the squared covariances in the column or row in which it is found.

Structural equation modeling methods

Confirmatory factor analysis within structural equation modeling (SEM), discussed in the Statistical Associates "Blue Book" on “Structural Equation Modeling,” is a common method of assessing convergent and discriminant validity. The researcher draws a model with the desired latent variables and assigns indicator variables to each depicted as ellipses (the latent variables) and arrows (connections to the indicator variables, showing convergent validity). Absence of connecting arrows from indicators for one latent variable to other latent variables indicate separation (discriminant validity). If goodness of fit measures for the measurement model in SEM are adequate, the researcher concludes that both convergent and divergent validity are upheld.

Nested models

A supplementary SEM-based approach to discriminant validity is to run the model unconstrained and also constraining the correlation between a pair of latent

variables (constructs) to 1.0. If the two models do not differ significantly on a likelihood ratio test of difference between models, the researcher fails to conclude that the constructs differ (see Bagozzi et al., 1991). In this procedure, if there are more than two constructs, one must employ a similar analysis on each pair of constructs, constraining the constructs to be perfectly correlated and then freeing the constraints. This method is considered more rigorous than either the SEM measurement model approach or the AVE method.

Example. In a study of industrial relations, Deery, Erwin, & Iverson (1999) wrote, "The discriminant validity was tested by calculating the difference between one model, which allowed the correlation between the constructs (with multiple indicators) to be constrained to unity (i.e., perfectly correlated), and another model, which allowed the correlations to be free. This was carried out for one pair of constructs at a time. For example, in testing organizational commitment and union loyalty, the chi-square difference test between the two models ($p<.001$) affirmed the discriminant validity of the constructs." See Deery, S., Erwin, P., & Iverson, R. (1999). Industrial relations climate, attendance behaviour, and the role of trade unions. *British Journal of Industrial Relations* 37(4): 533-558.

Criterion validity

Criterion validity, also called "concurrent validity" has to do with the correlation between measurement items on the one hand and known and accepted standard measures or criteria on the other. Such correlations are sometimes called “validity coefficients” (though that term has multiple meanings). Criterion validity may be thought of as proof of convergent validity: if items converge in meaning, the resulting scale should correlate with an established criterion variable.

Types of criterion validity

There are three types of criterion validity:

Correlation with an objective measure. Ideal criteria are direct, objective measures of what is being measured. For instance, a measure based on self-reported voting may be correlated with actual voting evidenced in voting records.

Correlation with an accepted measure. Where direct objective measures are unavailable, the criterion may be an accepted substitute measure. For instance, if

the researcher has developed a scale of "alienation", it should correlate highly with existing measures of the same construct accepted within the profession of psychology. Also, the correlations of the new construct with other related variables should be of the same direction and similar magnitude as for already validated scales of the alienation. Once demonstrated to be equivalent in terms of criterion validity, the researcher's new scale may be preferred if, for instance, it has fewer scale items and is thus more resistant to survey fatigue among respondents.

Correlation with other variables. In this type of criterion validity, the scale is correlated with other related variables to see if the proposed construct exhibits generally the expected direction and magnitude of correlation.

Predictive validity may be considered a variant on criterion validity, where the criterion is in the future. For instance, comparing if people who score high on an employment test also rate high on their subsequent job performance evaluations would be an effort to establish predictive validity because the criterion is not concurrent with the test.

Examples

Examples of criterion validity questions are: "Does a new measure of "alienation" exhibit the same general correlative behavior as established scales of social anomie?"; "Are people who score high on a proposed scale of "Republican identifiers" in fact more likely to vote Republican and/or be registered as Republican?"; and "Do people who score high on an employment test also rate high on current evaluations of actual job performance by their superiors?"

Example of criterion validity. An instrument measuring child-reported sexual abuse was analyzed for four different age groups of 103 Dutch children. There was only a weak correlation between the instrument and the outcome of the cases of child abuse (the objective criterion). The authors concluded that the results suggest that the instrument "cannot yet be used as a scientifically validated instrument for judging the truthfulness of allegations of child sexual abuse." Logic: A good instrument would identify valid cases of child abuse and for valid cases, the outcomes would tend to be different. See Lamers-Winkelman, F. & Heemstede, A.L. (1998). Statement validity analysis:

Its application to a sample of Dutch children who may have been sexually abused. *Journal of Aggression, Maltreatment & Trauma*, 2(2): 59-81.

Content validity

Overview

Also called *face validity*, content validity has to do with items seeming to measure what they claim to. In content validation the researcher is asking if the measures which operationalize concepts are ones which seem by common sense to have to do with the concept. Studies can be internally valid and statistically valid yet use measures lacking face validity. The items for a construct should measure the full domain of meaning implied by their label. Though derogated by some psychometricians as too subjective, failure of the researcher to establish credible content validity may easily lead to rejection of his or her findings. Use of panels of content experts, the Delphi method, or focus groups of representative subjects are ways in which content validity may be established, albeit using subjective judgments. (See also the separate Statistical Associates "Blue Book" volume on "The Delphi Method in Quantitative Research".)

There may be a *naming fallacy*: indicators may display construct validity yet the label attached to the concept may be inappropriate. In content validity the researcher asks if the labels attached to constructs are too broad in domain. For instance, in a small group study of corruption, "monetary incentives" may be used to induce participants to covertly break the rules of a game. The researcher may find "monetary incentives" do not lead to corruption, but what may be involved would be better labeled "small monetary incentives," whereas "large monetary incentives" may have a very different effect. Likewise, a scale may be labeled "liberalism" but the items may only deal with cultural issues like abortion and gay rights but lack any content on economic or environmental issues, and thus might be better labeled "cultural liberalism."

Example of content validity

Stratford and Kennedy (2004) noted this about a measure of functional status of lower extremities in a medical setting: "Clearly, a lower extremity functional status measure that does not overtly inquire about ambulation lacks content validity. For this reason, we caution against using the DISSIMILAR-8 as an outcome

measure for clinical trials and as the basis for decisions in clinical practice." See Stratford, P. W. & Kennedy, D. M. (2004). Does parallel item content on WOMAC's Pain and Function Subscales limit its ability to detect change in functional status? *BMC Musculoskeletal Disorders.* 2004; 5: 17.

Ecological validity

Ecological validity may be thought of as a special type of content validity. Not to be confused with the "ecological fallacy" discussed below under "statistical validity," ecological validity has to do with whether or not subjects are studied in their natural environments as compared, for example, to a lab setting. Removing subjects from their environment may lead subjects to display different behavior from their "true" behavior (ex., people may express different opinions about race in a lab setting compared to the opinions they express among friends). The greater the difference in cues and influences between the natural environment and the measurement setting, the greater the potential bias.

Four strategies are open to the researcher in dealing with ecological validity:

1. Showing ecological factors are not important
2. Taking measurement in the natural environment (where it is harder to control for external variables)
3. Replicating ecological factors in the measurement environment (ex., by priming questions in surveys)
4. Acknowledging that lab behavior cannot be extrapolated to behavior in the environment, but justifying this as a method of understanding behavior when removed from environmental influences

Internal validity

Internal validity has to do with defending against sources of bias arising in research design, typically by affecting the process being studied by introducing covert variables. When there is lack of internal validity, variables other than the independent variables being studied may be responsible for part or all of the observed effect on the dependent variables. See additional discussion in the Statistical Associates "Blue Book" volume on "Research Design".

Hawthorne effect (experimenter expectation)

Do the expectations or actions of the investigator contaminate the outcomes? This bias effect is named after famous studies at Western Electric's Hawthorn plant, where work productivity improvements were found to reflect researcher attention, not interventions like better lighting.

Mortality bias

Is there an attrition bias such that subjects later in the research process are no longer representative of the larger initial group?

Selection bias

How closely do the subjects approach constituting a random sample, in which every person in the population of interest has an equal chance of being selected? When multiple groups are being studied, there can be differential selection of the groups which can be associated with differential biases with regard to history, maturation, testing, mortality, regression, and instrumentation (that is, selection may combine differentially with other threats to validity mentioned on this page). The separate "blue book" on two-stage least squares regression contains a discussion of testing for selection bias.

Evaluation apprehension

Does the sponsorship, letter of entry, phrasing of the questions, or other steps taken by the researcher suffice to mitigate the natural apprehension people have about evaluations of their beliefs and activities, and diminish the tendency to give answers which are designed to make themselves "look good"?

Special problems involving control groups (social interaction threats to validity):

Compensatory equalization of treatments

Were those administering the setting pressured, or did they decide on their own, to compensate the control group's lack of the benefits of treatment by providing some other benefit for the control group? Parents may pressure school administrators, for instance, to provide alternative learning experiences to

compensate for their children in the control group not receiving the special test curriculum being studied in the experimental group.

Compensatory rivalry

Compensatory rivalry occurs when one group, typically the comparison or control group, learns of differences with the other group, typically the treatment group. Awareness of differences may create feelings of relative deprivation and may promote competitive attitudes which may bias the study. Compensatory rivalry has unpredictable outcomes and may lead on the one hand to feelings the group should work harder to compete with the other group, or that the group should give up trying in light of the known differences.

Resentful demoralization

Resentful demoralization is a type of compensatory rivalry where he comparison group learns of differences with the treatment group and becomes discouraged (and sometimes angry), leading to psychological or actual withdrawal from the study.

Treatment imitation or diffusion

This is also a type of control awareness invalidity, arising from the control group imitating the treatment or benefitting from information given to the treatment group and diffused to the control group.

Unintended treatments

The Hawthorne effect (see above) is an example, where the experimental group was also receiving the unmeasured "treatment" of researcher attention. However, either the experimental or control group may receive different experiences which constitute unmeasured variables.

Special problems of before-after studies and time series:

Cross-sectional limitations

Research designs which attempt to demonstrate change processes using point-in-time cross-sectional data often fail to provide the basis for confirmatory causal inference.

Instrumentation change

Variables may not be measured in the same way in the "after" study as in the "before" study. A subtle way for this to occur is when the observers/raters, through experience, become more adept at measurement.

History (intervening events)

Events not part of the study intervene between the before and after studies and may have an effect on the process under study. Did some historical event occur which would affect results? For instance, outbreak of a war often solidifies public opinion behind the Commander-in-Chief and could be expected to affect a study of causes of changes in presidential support in public opinion polls, even if the measurement items had nothing to do with foreign policy.

Maturation

Invalid inferences may be made when the maturation of the subjects between the before and after studies has an effect (ex., the effect of experience), but maturation has not been included as an explicit variable in the study.

Mortality

Subjects may drop out of the subject pool for a variety of reasons, making the "before" and "after" samples non-comparable.

Regression toward the mean

If subjects are chosen because they are above or below the mean, one would expect they will be closer to the mean on re-measurement, regardless of the

intervention. For instance, if subjects are sorted by skill and then administered a skill test, the high and low skill groups will probably be closer to the mean than otherwise expected.

Test experience

The before study may impact the after study in its own right, or multiple measurement of a concept may lead to familiarity with the items and hence a history or fatigue effect.

External validity

Overview

External validity has to do with possible bias in the process of generalizing conclusions from a sample to a population, to other subject populations, to other settings, and/or to other time periods. The questions raised are: "Are findings using the construct scale consistent across samples?"; "To what population does the researcher wish to generalize his/her conclusions?"; and "Is there something unique about the study sample's subjects, the place where they lived/worked, the setting in which they were involved, or the times of the study, which would prevent valid generalization?"

Naturally, when a sample is non-random in unknown ways, the likelihood of external validity is low, as in the case of convenience samples. All other things equal, different samples should generate similar relationships. When they do not, this indicates different samples are affected by significant unobserved variables which are not in the model and which differ in value across groups.

Example

In the NELS:88 study, indicators of student disability were obtained from students, parents, teachers, and school officials. Each indicator was worded differently, but all were intended to measure "student disability." Results of comparisons of these measures showed very little overlap (well under 50%) in the population of students identified as disabled by these separate sources. Logic: Different groups should have identified largely the same students as disabled, but they did not. This might mean all the similar instruments were invalid. But Rossi

and his associates chose to think it meant different wording had a large effect and the instruments measured different constructs even though that was the opposite of their intent. See Rossi, Robert, Jerald Herting, and Jean Wolman. 1997. *Profiles of Students With Disabilities as Identified in NELS:88* (NCES 97-254). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Statistical validity

Statistical validity has to do with basing conclusions on proper use of statistics. Violation of statistical assumptions is treated elsewhere, in the discussion of each specific statistical procedure. In addition, the following general questions may be asked of any study.

Reliability

Has the research established the statistical reliability of his/her measures? A measure is reliable if measurement of the same phenomena at different times and places yields the same measurement. Reliability is discussed below.

Type I errors and statistical significance

A Type I error occurs when the researcher thinks there is a relationship but there really isn't (a "false positive"). If the researcher rejects the null hypothesis because p<=.05, leading to the conclusion there *is* a relationship, ask these questions:

- Are data from an enumeration rather than a sample? If so, significance tests do not apply as it is meaningless to ask, as significance tests do, if taking another sample would generate results as strong or stronger than those observed.
- Are data from a non-random sample? If so, significance coefficients are in error to an unknown degree. While bootstrap significance tests may be applied, bootstrap significance has a different, limited interpretation having to do with the chances of obtaining results as strong or stronger by chance of resampling (taking another sample from the data at hand) rather than by chance of sampling (taking another random sample from the population).

- Are data from a randomized experiment? If so, significance has to do with the chances of obtaining results as strong or stronger by chance of re-randomization of subjects (re-randomizing the subjects) rather than by chance of sampling (taking another random sample from the population). Randomized experiments asymptotically control for unobserved variables but do not control for sampling bias in the subject pool (ex., if subjects are entirely of one race it is impossible to generalize validly to another race).
- If data are from a random sample, is significance established to be of an appropriate level? In social science, this level is usually .05 for confirmatory analysis but often only .10 for exploratory analysis.
- Are significance tests applied to à *priori* hypotheses, or is a shotgun approach used in which large numbers of relationships are examined, looking à *posteriori* for significant ones? If the latter, note that 1 table or relationship in 20 will be found to be statistically significant just by chance alone, by definition of .05 significance. Multiple à *posteriori* tests require a higher operational alpha significance level to achieve the same nominal alpha level. This is achieved by using Bonferroni or other adjustments, discussed separately in the "blue book" on significance tests and in "blue books" on certain individual statistical procedures.

See further discussion in the Statistical Associates “Blue Book” volume on “Significance Testing”.

Type II Errors and Statistical Power

A Type II error is when the researcher thinks there is no relationship, but there really is (“false negatives”). If the researcher has accepted the null hypothesis because p>.05, leading to the conclusion there is no relationship, ask these questions:

- Has the researcher used statistical procedures of adequate power? If power is equal to or greater than .80, a no-relationship conclusion is considered valid. Significance only allows failure to reject the null hypothesis, which is different from accepting the null hypothesis (that requires power >= .80).
- Does failure to reject or accepting the null hypothesis merely reflect small sample size?

Interaction and non-linearity

Another issue of statistical validity is whether the researcher has taken possible interaction effects and nonlinear effects into account. Likewise, is there interaction among multiple treatments?

Causal ambiguity

Has the researcher misinterpreted the causal direction of relationships, particularly in correlative studies?

Fallacies of aggregation

Discussed by the statistician G. Udny Yule early in the 20th century, called the "ecological fallacy" by Robinson (1950), and "Simpson's Paradox" based on 1951 work by Edward Simpson, the ecological fallacy is assuming that individual-level correlations are the same as aggregate-level correlations, or vice versa. Robinson showed that individual level correlations may be larger, smaller, or even reverse in sign compared to aggregate level correlations. For instance, at the state level in the USA there is a strong correlation of race and illiteracy but this largely disappears at the individual level. The reason is that many African-Americans live in the South, which also has high illiteracy for whites as well. More generally, what is true at one level of aggregation (ex., states) need not be true at another level (ex., individuals or nations). Various methods have been proposed for dealing with ecological inference discussed below, but there is no simple solution.

Strategies for dealing with fallacies of aggregation

There is no good "solution" to the problems posed by the ecological fallacy. The monograph by Langbein and Lichtman (1978) addressed this long ago. They debunked a number of purported procedures, such as "homogenous grouping" (restricting analysis to aggregate units which do not vary on variables of interest, as by looking at counties which are very high in Catholicism and counties very high in Protestantism to assess the individual-level voting behaviors or Catholics and Protestants), which they note has "disabling defects" (p. 39). Langbein and Lichtman found the best of bad alternatives to be "ecological regression" (pp. 50-60), a technique which involves using regression to model the effects of grouping.

Ecological regression, however, has often been criticized as vulnerable to statistical bias. Ecological regression fails whenever factors affecting the dependent variable co-vary with the independent variables. Epstein and O'Halloran (1997) give a good critique of ecological regression, noting, "For instance," (in a study in which the dependent is votes for the minority candidate),"say that voter registration rates are not constant across districts, but rather minority registration rates increase as the percent of Black voters in a district rises. In such a case, ecological regression would attribute the rise in the number of votes for the minority candidate to polarized voting, when in fact it is partly due to increased minority registration." (p. 4). Epstein and O'Halloran propose logit and probit techniques instead of ecological regression. See Achen and Shively (1995) and King (1997).

Today the accepted approach is to have data at both levels and to model higher-level effects on lower-level relationships using multilevel modeling, also known as linear mixed modeling or hierarchical linear modeling.

Validity Checklist

1. If you have a construct which is based on a set of indicator items, have you shown convergent validity (ex., that Cronbach's alpha > .7, or show items are an ordered Guttman or other scale; further discussion in section on multidimensional scaling)

2. If you have shown convergent validity and if you have more than one construct, have you also shown divergent validity? If you cannot show divergent validity, the two predictor constructs may be multicollinear or even reflect a single common factor. If one construct is a predictor and the other a dependent and you cannot show divergent validity, results may be specious due to definitional overlap. Factor analysis is one way to establish divergent validity, and SEM a different and better way.

3. If there is a known, accepted measures of a construct, have you shown your construct is highly correlated with it? This is criterion validity. For instance, perhaps you have a 5-item scale which can be used in lieu of an accepted 20-item scale.

4. Do you restrict your conclusions to what your sample warrants? Nothing beats having a random sample of the population to which you wish to generalize. This is external validity, for which a subtype is ecological validity, which is whether you studied subjects in their natural setting, not some unrealistic setting (ex., a lab). You need to show there is no apparent selection bias. For instance, selecting extreme cases is subject to regression toward the mean.

5. Your items should have content validity (face validity): knowledgeable observers should agree that the items seem to measure what you say. You could have a panel of experts make a judgment on content validity, or just have your pretest population do so in debriefing.

6. Avoid interviewer effects (avoid Hawthorne effects), which are a type of placebo effect (subjects respond to attention).

7. Avoid multilevel fallacies. The ecological fallacy is assuming that what is true at the individual level is true at the group level. Moreover, if there are multilevel effects at a grouping level (ex., teacher classroom and school effects on individual-level student performance), some form of linear mixed modeling is required.

8. Statistical validity: did you meet the assumptions of the procedure you used?

9. Avoid common method bias such as the yea-saying bias in survey research. Multi-method, multi-trait approaches are idea but not common.

RELIABILITY ANALYSIS OVERVIEW

Researchers must demonstrate instruments are reliable since without reliability, research results using the instrument are not replicable, and replicability is fundamental to the scientific method. Reliability is the correlation of an item, scale, or instrument with a hypothetical one which truly measures what it is supposed to. Since the true instrument is not available, reliability is estimated in one of four ways:

- *Internal consistency*: Estimation based on the correlation among the variables comprising the set (typically, Cronbach's alpha)
- *Split-half reliability*: Estimation based on the correlation of two equivalent forms of the scale (typically, the Spearman-Brown coefficient)
- *Test-retest reliability*: Estimation based on the correlation between two (or more) administrations of the same item, scale, or instrument for different times, locations, or populations, when the two administrations do not differ on other relevant variables (typically, the Spearman Brown coefficient)
- *Inter-rater reliability*: Estimation based on the correlation of scores between/among two or more raters who rate the same item, scale, or instrument (typically, intraclass correlation, of which there are six types discussed below).

These four reliability estimation methods are not necessarily mutually exclusive, nor need they lead to the same results. All reliability coefficients are forms of correlation coefficients, but there are multiple types discussed below, representing different meanings of reliability and more than one might be used in single research setting.

Reliability: Overview

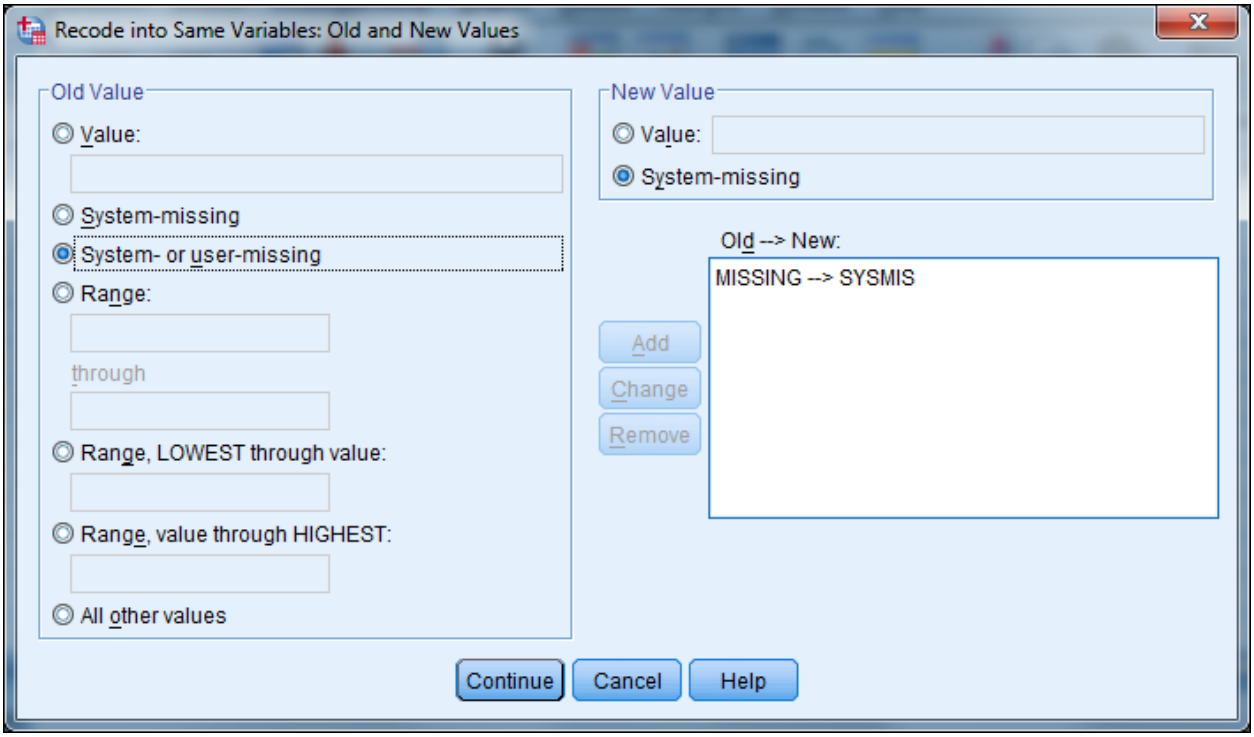
Data

In the examples in this section, the "GSS93 subset.sav" dataset is used. This dataset is provided in the "Samples" directory when SPSS is installed. In the section on Cohen's kappa, the data shown below are small enough to type in. In the ICC section, the dataset used is "tv-survey.sav", also from the SPSS "Samples"

directory. A Google search for '"GSS93 subset.sav" download' or '"tv-survey.sav" download' will reveal hundreds of hits.)

SPSS will save a dataset to SAS .sas7bdat format and to Stata .dta format, or alternatively both SAS and Stata can import SPSS .sav files. Before saving from SPSS to SAS or Stata formats, however, it is important to recode all user-missing values to be system-missing, otherwise user-missing codes may be used as real values. Also note that SAS does not allow dashes in file names, so the SAS version must start with "tv_survey", not "tv-survey". Stata will allow dashes.

To write to Stata's .dta data format from SPSS, the researcher must first replace all user-missing values with system-missing values because Stata does not recognize SPSS user-missing codes (typically, 0, 8, 9, 99, etc.) the user may have defined. To do this in SPSS select Transform > Recode into Same Variables; then select all variables and click the "Old and new values" button. In the dialog shown below, replace all "system- or user-missing" with "System missing" (check as below, click Add, Continue, OK).



After this replacement, then simply select File > Save as, and then select the Stata flavor wanted, as shown in the figure below. Then in Stata, select File > Open, browse to the directory where the file was saved, and select it (here, GSS93 subset.dta).

Measurement

Scores

Scores are the subject's responses to items on an instrument (ex., a mail questionnaire). *Observed scores* may be broken down into two components: the *true score* (commonly labeled tau) plus the *error score*. The error score, in turn, can be broken down into *systematic error* (non-random error reflecting some systematic bias, as due, for instance, to the methodology used -- hence also called *method error*) and *random error* (due to random traits of the subjects -- hence also called *trait error*). The smaller the error component in relation to the true score component, the higher the reliability of an item, which is the ratio of the true score to the total (true + error) score.

Number of scale items

Note that the larger the number of items added together in a scale, the less random error matters as it will be self-cancelling (think of weighing a subject on 100 different weight scales and averaging rather than on using just one scale), and therefore some reliability coefficients (such as Cronbach's alpha) also compute higher reliability when the number of scale items is higher.

Triangulation

Triangulation is the attempt to increase reliability by reducing systematic (method) error, through a strategy in which the researcher employs multiple methods of measurement (ex., survey, observation, archival data). If the alternative methods do not share the same source of systematic error, examination of data from the alternative methods gives insight into how individual scores may be adjusted to come closer to reflecting true scores, thereby increasing reliability.

Calibration

Calibration is the attempt to increase reliability by increasing homogeneity of ratings through feedback to the raters, when multiple raters are used. Raters meet in calibration meetings to discuss items on which they have disagreed,

typically during pretesting of the instrument. The raters seek to reach consensus on rules for rating items (ex., defining the meaning of a "3" for an item dealing with job satisfaction). Calibration meetings should not involve discussion of expected outcomes of the study, as this would introduce bias and undermine validity.

Models

In SPSS

Statistical models for reliability analysis supported by SPSS (Analyze > Scale > Reliability Analysis) under the "Models" button of the reliability dialog are listed below. Subsequent sections of this volume treat how to obtain the corresponding coefficients in SPSS, SAS, and Stata:

- *Alpha (Cronbach).* This models internal consistency based on average correlation among items.
- *Split-half.* This model is based on the correlation between the parts of a scale which is split into two forms.
- *Guttman.* This is an alternative split-half model which computes Guttman's lower bounds for true reliability, discussed below.
- *Parallel.* This method uses maximum likelihood to test if all items have equal variances and error variances. Cronbach's alpha is the maximum likelihood estimate of the reliability coefficient when the parallel model is assumed to be true (SPSS, 1988: 873). If the chi-square goodness of fit significance for the parallel model is <=.05, the researcher rejects the null hypothesis that the items have equal variances and error variances in the population. This implies Cronbach's alpha would be different between original and standardized data.
- *Strict parallel.* This method also uses maximum likelihood to test for equal variances, equal error variances, and equal population means across items.

In SAS

PROC RELIABILITY in SAS does not correspond to the SPSS Reliability module. Rather, PROC RELIABILITY provides a variety of reliability coefficients associated with survival analysis with uncensored, right-censored, interval-censored, and

arbitrarily censored data. In SAS, the following commands are those supporting types of reliability discussed in this volume:

- PROC CORR followed by the ALPHA parameter computes Cronbach's alpha.
- PROC FREQ computes Cohen's kappa (simple and weighted) when the AGREE option is specified in the TABLES statement. If you also specify the KAPPA option in the TEST statement, then PROC FREQ computes the asymptotic test of the hypothesis that simple kappa equals zero. Similarly, if you specify the WTKAP option in the TEST statement, PROC FREQ computes the asymptotic test for weighted kappa.

To this author's knowledge, there is SAS does not directly support Spearman-Brown split half reliability. Of course, an instrument can be split manually and the two parts correlated with PROC CORR. As far as this author can determine, Guttman's lower bounds also is not supported in SAS. SAS supports the usual type of intraclass correlation only through third-party SAS macros (see below).

In Stata

In Stata, the following commands are those supporting types of reliability discussed in this volume:

- alpha: Compute inter-item correlations (covariances) and Cronbach's alpha, including item-test and item-rest correlations. This corresponds to the menu selections Statistics > Multivariate analysis > Cronbach's alpha.
- 1oneway: Calculates the intraclass correlation coefficient for one-way analysis of variance models.
- icc: Calculates the intraclass correlation coefficient for one-way random-effects models, for two-way random-effects models. And two-way mixed-effects models.
- kappa: Computes Cohen's kappa for interrater agreement.
- correlate: Calculates correlations and covariances of variables or coefficients.

Stata does not support Spearman-Brown split-half reliability. There is a third-party .ado file called sbrowni, which computes the Spearman-Brown correction for test length described below but does not compute the Spearman-Brown coefficient. In addition, there is also a third-party .ado file called sbri (.ssc install sbri). This

program calculates a Spearman-Brown reliability coefficient of a different type. Rather than comparing two or more batteries of test items for the same construct, it must be used as a postestimation command after xtreg regression with the mle (maximum-likelihood random-effects estimator) option. It then compares xtreg results across groups formed by a categorical variable (ex., across automobile manufacturers when predicting miles per gallon). As this is a different context for the Spearman-Brown coefficient, it is not discussed here.

As far as this author can determine, Guttman's lower bounds also is not supported in Stata.

Internal consistency reliability

Cronbach's alpha

Overview

Cronbach's alpha is the most commonly used internal consistency reliability coefficient. It is used as a convergent validity coefficient. Alpha equals zero when the true score is not measured at all and there is only an error component. Alpha equals 1.0 when all items measure only the true score and there is no error component. Since the "true score" is unknown, like all reliability coefficients, alpha is an estimate.

Interpretation

Cronbach's alpha can be interpreted as the percent of variance the observed scale would explain in the hypothetical true scale composed of all possible items in the universe. Alternatively, it can be interpreted as the correlation of the observed scale with all possible other scales measuring the same thing and using the same number of items.

Cut-off criteria

By convention, a lenient cut-off of .60 is common in exploratory research. Alpha should be at least .70 or higher to retain an item in an "adequate" scale for confirmatory research and should be .80 or higher for a "good scale" in

confirmatory research. Since reliability is presumed the same using any reliability coefficient, these same cut-offs apply to other reliability coefficients as well.

Formula

Standardized alpha is a ratio. The numerator is the number of items times the average of the covariances of all pairs of items. The denominator is the average variance of the items plus the quantity N-1 times the average covariance, where N is the number of items.

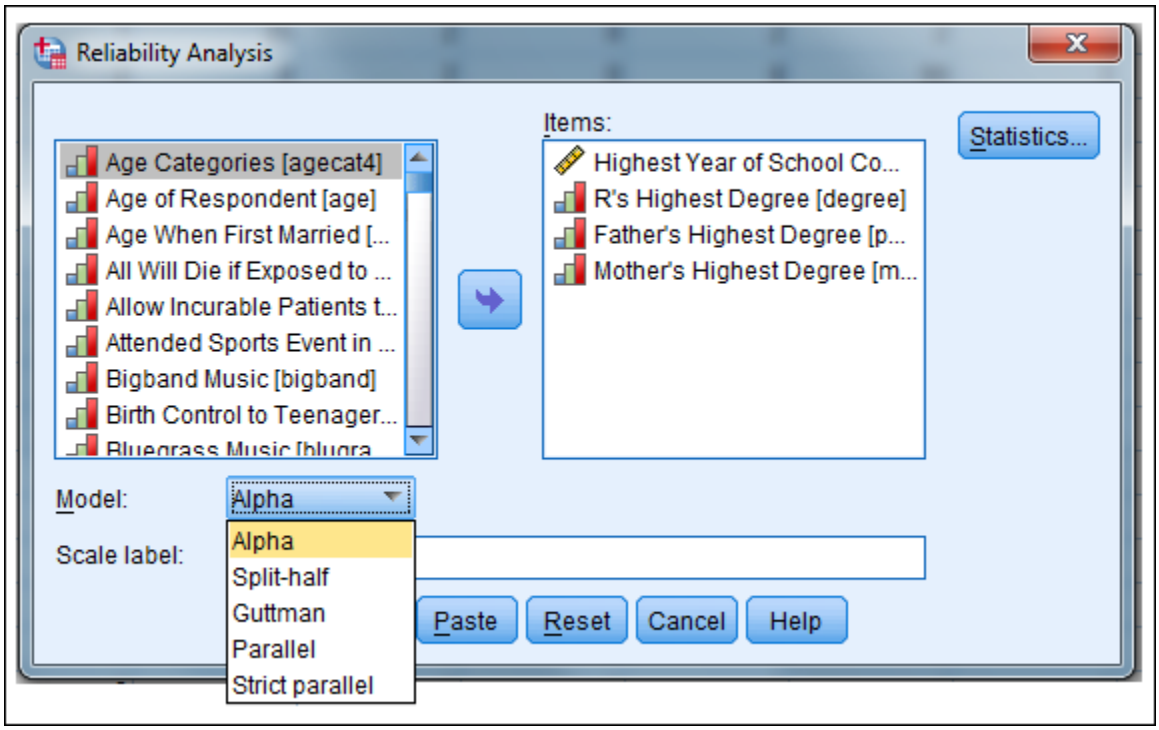
Number of items

Note that Cronbach's alpha increases as the number of items in the scale increases, even controlling for the same level of average intercorrelation of items. This assumes, of course, that the added items are not bad items compared to the existing set. Increasing the number of items can be a way to push alpha to an acceptable level. This reflects the assumption that scales and instruments with a greater number of items are more reliable. It also means that comparison of alpha levels between scales with differing numbers of items is not appropriate. However, Stata supports the Spearman-Brown reliability correction for test length, discussed below.

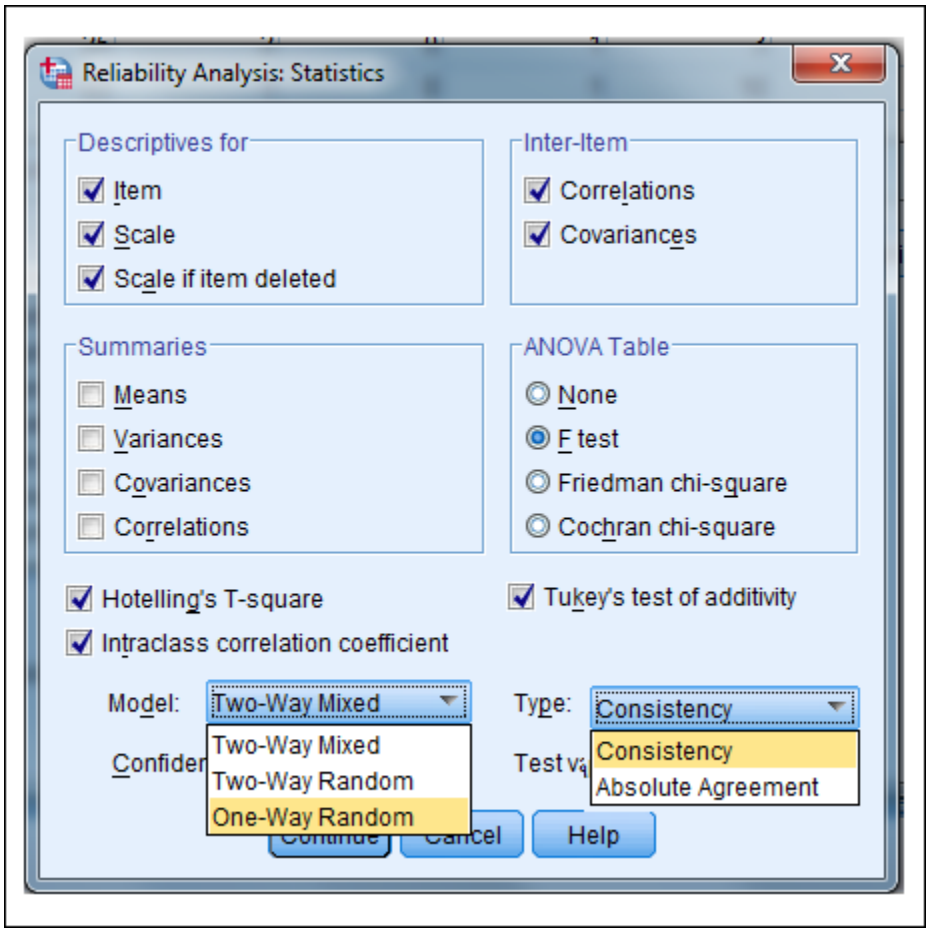
Cronbach's alpha in SPSS

SPSS user interface

After selecting Analyze > Scale > Reliability Analysis from the menus, the dialog shown below appears. The “Model:” drop-down menu enables the researcher to select among a variety of model types, though in this subsection “Alpha” is selected (the default), to compute Cronbach’s alpha. The figure also shows that four education items have been selected for evaluation, following “Example 1” below.



Clicking the "Statistics" button opens a dialog in which output may be selected, shown below. The default is window shows nothing checked but here is shown for the example. The figure also shows the drop-down menu for various models for the intraclass correlation coefficient, discussed further below.



SPSS statistical output

The "Reliability Statistic" table

This table, which is default SPSS output, prints the usual Cronbach's alpha coefficient in the first column, here .707. As this is greater than .70, the scale is judged adequate for confirmatory research.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.707	.798	4

Note that occasionally alpha will be negative. A negative Cronbach's alpha indicates inconsistent coding or a mixture of items which measure different

dimensions, leading to negative inter-item correlations. This is discussed in the FAQ section below.

"Cronbach's alpha based on standardized items", better known as standardized item alpha, is higher (.798) in the figure above. Standardized item alpha is the average inter-item correlation when item variances are equal. (It is also called the *Spearman-Brown stepped-up reliability coefficient* or simply the "Spearman-Brown Coefficient," but these terms should not be confused with the better-known Spearman-Brown split-half reliability coefficient discussed below.) The difference between Cronbach's alpha and standardized item alpha is a measure of the dissimilarity of variances among items in the set.

The "Item Statistics" table shows the mean and standard deviation of the items. Below it is shown that highest year of school completed is of a higher magnitude and greater variance than the other items, a fact which might lead the researcher to prefer standardized item alpha over alpha.

Item Statistics

	Mean	Std. Deviation	N
Highest Year of School Completed	13.32	3.045	1124
R's Highest Degree	1.54	1.189	1124
Father's Highest Degree	.96	1.204	1124
Mother's Highest Degree	.85	.938	1124

In a second use, standardized item alpha can be used to estimate the change in reliability as the number of items in an instrument or scale varies.

$r_{S2} = (N * r_{ave})/[1 + (N-1) * r_{ave})]$
where

r_{S2} = standardized item alpha
r_{ave} = the average of inter-item correlations
N = total number of items

"Hotelling's T-Squared Test" table

If requested, SPSS generates the Hotelling T-Squared test statistic. As shown in the figure below, it is significant (.000) for these data. A finding of significance in this context means that the researcher rejects the null hypothesis that the means

of the items are equal. This was already evident simply by examining the "Item Statistics" table shown above and it has the same implication: standardized item alpha may be preferred over alpha due to differences in the magnitudes of the items. Hotelling's test assumes multivariate normality of items. See further discussion in the "Assumptions" section below.

Hotelling's T-Squared Test

Hotelling's T-Squared	F	df1	df2	Sig
39068.394	12999.605	3	1121	.000

The "Item-Total Statistics" table and alpha if item deleted

The item-total correlation is the Pearsonian correlation of the item with the total of scores on all other items. A low item-total correlation means the item is little correlated with the overall scale (ex., < .3 for large samples or not significant for small samples) and the researcher should consider dropping it. A negative correlation indicates the need to recode the item in the opposite direction. The reliability analysis should be re-run if an item is dropped or recoded. Note a scale with an acceptable Cronbach's alpha may still have one or more items with low item-total correlations.

"Cronbach's Alpha if Item Deleted" in the last column of the figure below is the estimated value of alpha if the given item were removed from the model. It is not default output but rather the researcher must check the appropriate box in the "Statistic" button dialog shown above.

This table here shows that dropping any item will lower alpha. This is especially true of "RS Highest Degree", which in the "Inter-Item Correlations" table further below is shown correlated at the .872 level with "Highest Year of School Completed."

Item-Total Statistics

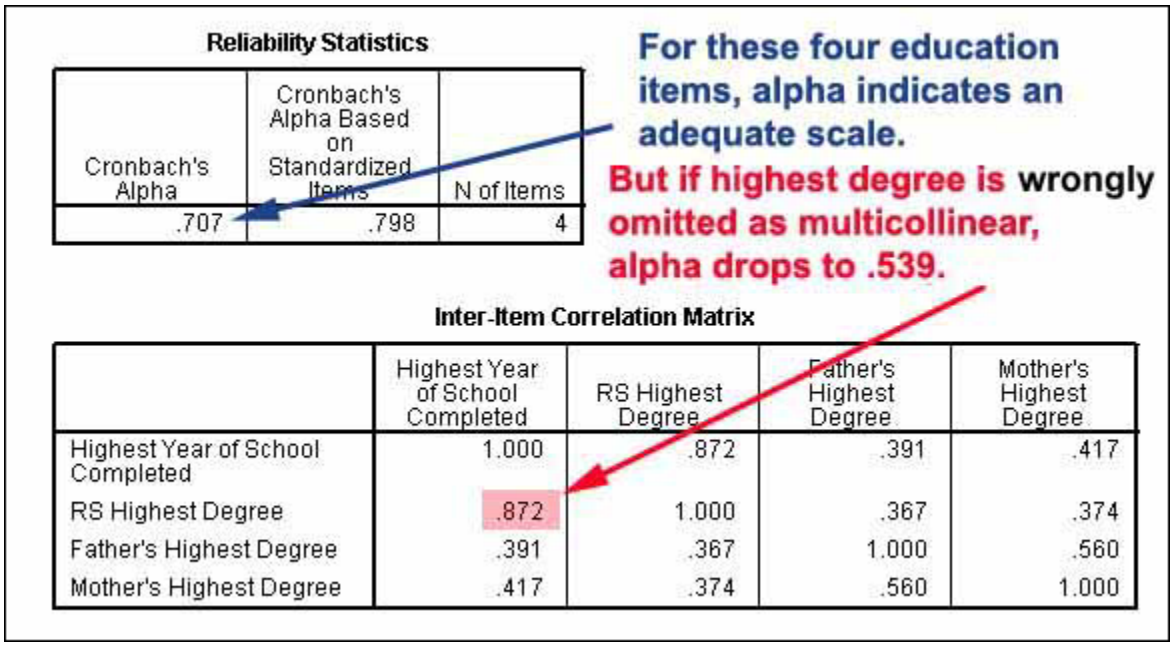
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Highest Year of School Completed	3.35	6.894	.723	.770	.686
R's Highest Degree	15.13	18.118	.810	.761	.539
Father's Highest Degree	15.70	21.098	.469	.345	.677
Mother's Highest Degree	15.81	22.368	.505	.360	.686

If "alpha if deleted" is lower for all items than for the computed overall alpha, then no items need be dropped from the scale. Also, if "alpha if deleted" is above the cut-off level acceptable to the researcher (ex., above .8), the researcher may wish to drop that item anyway simply in order to have fewer items in the instrument. Note, however, that when an item has high random error it is possible that it would be removed on this basis when, in fact, it does measure the same construct.

The squared multiple correlation in the next-to-last column in the figure above is the R^2 for an item when it is predicted from all other items in the scale. The larger the R^2, the more the item is contributing to internal consistency. The lower the R^2, the more the researcher should consider dropping it. Note the R^2 of some items may be low even on a scale which has an acceptable Cronbach's alpha overall.

The "Inter-Item Correlation Matrix" table

The "Inter-Item Correlation Matrix" table throws additional light on the items which make up the scale. (Note there is also an "Inter-Item Covariance Matrix" table.) The highest correlation in the set is the correlation of highest year of school completed with respondent's highest degree (.872). The lowest correlation is between respondent's highest degree and father's highest degree (.367).



Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.707	.798	4

Inter-Item Correlation Matrix

	Highest Year of School Completed	RS Highest Degree	Father's Highest Degree	Mother's Highest Degree
Highest Year of School Completed	1.000	.872	.391	.417
RS Highest Degree	.872	1.000	.367	.374
Father's Highest Degree	.391	.367	1.000	.560
Mother's Highest Degree	.417	.374	.560	1.000

Some authors state that it is desirable that inter-item correlations be not very strong (ex., Meyers et al., 2012: 725). Taken out of context, this is misleading as it would suggest that lower convergent validity is desirable in a scale. Higher inter-item correlations will increase alpha. If two items are so highly correlated as to be multicollinear, then one may be dropped as redundant, making for a shorter scale. This is desirable as long as, unlike in the example above, alpha does not drop below a critical level. It is also true that one may validate a scale as adequate even with quite modest inter-item correlations, showing Cronbach's alpha to be a fairly lenient criterion for validating convergent validity and the unidimensionality of a scale. That is all the more reason why the alpha cut-off for a "good scale" should be .80, not some lower figure.

Correlations over .80 may signal multicollinearity, which in turn might wrongly lead the researcher in this example to drop "RS Highest Degree" from the scale, in turn leading to the conclusion that the remaining three items do not constitute a scale suitable even for exploratory purposes (since the resulting alpha if deleted is .539, which is below the .60 cutoff level even for exploratory purposes). However, high correlation of the constituent items of a scale typically is not considered multicollinearity because the scale score, not the separate item scores, will appear in the regression (or other) analysis which uses the scale.

The "ANOVA with Tukey's Test of Nonadditivity" table

If requested, SPSS generates an ANOVA table showing Tukey's test of non-additivity. In the "Residual Nonadditivity" row, the F test shows significant additivity (p = .000). In a good model, this test is non-significant. The footnote to the table shows the power to which each score may be raised to correct for this.

ANOVA with Tukey's Test for Nonadditivity

			Sum of Squares	df	Mean Square	F	Sig
Between People			7785.301	1123	6.933		
Within People	Between Items		125804.641	3	41934.880	20674.086	.000
	Residual	Nonadditivity	3616.005[a]	1	3616.005	3785.022	.000
		Balance	3217.604	3368	.955		
		Total	6833.609	3369	2.028		
	Total		132638.250	3372	39.335		
Total			140423.551	4495	31.240		

Grand Mean = 4.17

a. Tukey's estimate of power to which observations must be raised to achieve additivity = .463.

Non-additivity means that there is interaction of the scale items with subjects as a factor. Put another way, non-additivity means there is a multiplicative interaction between the subjects and the items. After transformation by raising each score to a power of .463, the items may then be added to form a scale. Without this, the item scores are non-additive. See further discussion of additivity and Tukey's test in the "Assumptions" section below

KR20 and KR21

The Kuder-Richardson (KR20) *coefficient* is the same as Cronbach's alpha when items are dichotomous. Kuder-Richardson KR21 is a variant used when items are dichotomous and similar in difficulty.

$KR20 = [n/(n - 1)]*[1 - (\sum pq)/v]$

$KR21 = [n/(n-1)]*[1 - (m*(n - m)/(n*v))]$

Where

n = number of items

∑pq is the sum of the product of p and q, where p is the proportion of subjects passing the item and q is the proportion failing
m = mean score on the scale
v = variance of the scale

Cronbach's alpha in SAS

SAS syntax

In SAS Cronbach's alpha is part of PROC CORR. The syntax below loads the SPSS sample dataset "GSS93 subset.sav" and executes PROC CORR, asking for Cronbach's alpha output. Statements beginning with asterisks are comments, not executed by SAS.

```
* PROC IMPORT imports the data. Actual directory paths belong
within the DATAFILE quotes but are omitted here;
PROC IMPORT OUT= WORK.Cronbach
            DATAFILE= "GSS93 subset.sav"
            DBMS=SPSS REPLACE;
run;

* PROC CORR followed by the alpha parameter causes
Cronbach alpha output to be computed;
* The nomiss parameter drops cases with missing values;
* The nocorr parameter suppresses correlation output;
* The DATA statement refers back to the user-set name for
the work file from the IMPORT statement above;
PROC CORR alpha nomiss nocorr DATA=Cronbach;

* The var statement lists the variables to be used;
VAR educ degree padeg madeg;
run;
```

SAS output

SAS output differs in format but gives approximately equal Cronbach alpha values (small differences due to rounding and algorithms).

The CORR Procedure

4 Variables: educ degree padeg madeg

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
educ	1124	13.31673	3.04540	14968	0	20.00000
degree	1124	1.53559	1.18899	1726	0	4.00000
padeg	1124	0.95996	1.20373	1079	0	4.00000
madeg	1124	0.85053	0.93839	956.00000	0	4.00000

Simple Statistics

Variable	Label
educ	Highest Year of School Completed
degree	R's Highest Degree
padeg	Father's Highest Degree
madeg	Mother's Highest Degree

Cronbach Coefficient Alpha

Variables	Alpha
Raw	0.707414
Standardized	0.797945

Cronbach Coefficient Alpha with Deleted Variable

Deleted Variable	Raw Variables Correlation with Total	Raw Variables Alpha	Standardized Variables Correlation with Total	Standardized Variables Alpha	Label
educ	0.722924	0.685586	0.709522	0.696821	Highest Year of School Completed
degree	0.809961	0.539309	0.673547	0.715411	R's Highest Degree
padeg	0.468769	0.677488	0.524145	0.788558	Father's Highest Degree
madeg	0.504908	0.686086	0.539931	0.781132	Mother's Highest Degree

Cronbach's alpha in Stata

Stata syntax

The same data file and items as in SPSS and SAS examples, except in Stata .dta format, is opened by the .use command:

```
. use "C:\Data\GSS93 subset.dta", clear
```

Cronbach's alpha is computed with the .alpha command. Deletion of cases with missing values is not the default, hence the "casewise" option following the comma. The "detail" option generates a listing of individual interitem correlations and covariances. The "item" option generates a listing of item-test and item-rest correlations.

```
. alpha educ degree padeg madeg, casewise detail item
```

To obtain standardized item alpha, simply add the std option.

```
. alpha educ degree padeg madeg, casewise detail item std
```

Additional options not illustrated below include generate (save the newly generated scale in a new variable in the dataset), label (display variable labels), and reverse (reverse the signs of specified variables before calculating alpha).

Stata output

Stata produces the virtually identical estimates of alpha and standardized item alpha as do SPSS and SAS. Below, only the output for standardized data is shown (using the std option). Output for unstandardized data would be parallel except output is for covariances rather than correlations, and the "alpha" column is Cronbach's alpha rather than standardized item alpha.

Interpretation of most of the output parallels discussion above in the SPSS section on Cronbach's alpha. Novel (and useful) in Stata output is the "item-rest correlation" column, showing the correlation of the row item with a scale composed of the other items. Tukey's test of non-additivity and Hotelling's coefficient, discussed above in the SPSS section, are not supported within the Stata alpha command.

```
. alpha educ degree padeg madeg, casewise detail item std

Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
educ	1124	+	0.8489	0.7095	0.4338	0.6968
degree	1124	+	0.8279	0.6735	0.4559	0.7154
padeg	1124	+	0.7345	0.5241	0.5542	0.7886
madeg	1124	+	0.7448	0.5399	0.5433	0.7811
Test scale					0.4968	0.7979

```
Interitem correlations (obs=1124 in all pairs)
```

	educ	degree	padeg	madeg
educ	1.0000			
degree	0.8716	1.0000		
padeg	0.3911	0.3671	1.0000	
madeg	0.4167	0.3743	0.5599	1.0000

Spearman-Brown reliability correction for test length

A third party has made available the sbrowni.ado file, which performs a Spearman-Brown reliability correction for test length. It is installed with the command `.ssc install sbrowni`. As such the sbrowni program can adjust the Cronbach's alpha reliability coefficient or the Kuder-Richardson reliability coefficient (which is alpha for dichotomous items

The sbrowni program estimates how much Cronbach's alpha (or the Kuder-Richardson coefficient, which is alpha for dichotomous items) would increase if the number of items were increased to a certain number; and it estimates the number of items required to obtain a particular reliability of a specified level. This test is of interest because Cronbach's alpha depends in part on number of items in a scale.

The sbrowni program does not require a dataset to be in use. Rather, the researcher supplies previously-computed reliability coefficients and number of items in a scale. There are two uses, described below.

1. To estimate how much alpha would increase if the number of items were increased to a certain number. The syntax is

   ```
   .sbrowni #rel0 #count0 #count1
   ```

   where #rel10 is the previously-computed reliability, #count0 is the existing count of items in the scale, and #count1 is the proposed higher count. Thus the command

   ```
   .sbrowni .70 20 60
   ```

   returns the alpha level if a scale of 20 items with a present Cronbach's alpha of .70 were increased to a 60 item scale. Output is shown below.

```
. sbrowni .70 20 60

Spearman-Brown reliability correction for test length
-----------------------------------------------------

Initial reliability     = 0.700
Initial number of items =  20
Increased test length   =  60

Estimated reliability   = 0.875
```

2. The sbrowni program will also estimate the number of items required to obtain a particular reliability. The syntax is

   ```
   .sbrowni #rel0 #count0 #rel1
   ```

   where #rel10 is the previously-computed reliability, #count0 is the existing count of items in the scale, and #rel1 is the desired higher reliability. Thus the command

   ```
   .sbrowni .50 20 .80
   ```

   returns the number of items needed if a scale of 20 items with a present Cronbach's alpha of .50 were to obtain an alpha of .80. Output is shown below.

```
. sbrowni .50 20 .80

Spearman-Brown reliability correction for test length
-----------------------------------------------------

Initial reliability     = 0.500
Initial number of items =  20
Desired reliability     = 0.800

Estimated test length   =   80
```

Other internal consistency reliability measures

Ordinal coefficient alpha and ordinal coefficient theta

Simulation studies by Zumbo, Gaderman, & Zeisser (2007) have shown that for binary and ordinal data, Cronbach's alpha underestimates the true/theoretical reliability in three circumstances:

- The fewer the response options, the greater the downward bias, especially for fewer than 5 options
- The lower the true reliability, the more the downward bias, especially below .80.
- The greater the skew, the greater the downward bias..

Under these circumstances, Cronbach's alpha is a conservative estimate or lower bound for reliability. These findings suggest that if a scale of binary or ordinal items is judged reliable by Cronbach's alpha, it is more reliable than alpha indicates. If it is judged unreliable, the researcher may be making a Type II error.

Where Cronbach's alpha is based on Pearson correlation, which assumes interval data, Zumbo and his colleagues showed that it is possible to compute an ordinal version of alpha based on polychoric correlation, which assumes only ordinal data. Specifically, a matrix of polychoric correlations can be input into factor analysis (principal axis factoring) and then using formulas described below, ordinal coefficient alpha may be computed. In a second variant, the polychoric correlation matrix can be input into principal components analysis and then using a formula describe below, ordinal coefficient theta can be calculated.

Both ordinal coefficient alpha and ordinal coefficient theta are estimates of reliability. As such they have the same rule of thumb cut-off criteria: .80 or higher for confirmatory research involving small effects or larger; .70 or higher adequate for confirmatory research involving medium to strong effects; and .60 or higher for exploratory research.

Principal factor analysis (PFA) is used in causal modeling, as in structural equation modeling, so ordinal coefficient alpha would be preferred for most social science applications. Principal components analysis (PCA) is used when the research purpose is data reduction, possibly for exploratory purposes. For a principal components model, ordinal coefficient theta would be appropriate. (PFA and PCA are compared in greater depth in the separate Statistical Associates "Blue Book" volume on "Factor Analysis.")

In light of the above, Liu, Wu, & Zumbo (2010: 21) state, "Even without outliers, Cronbach's alpha may not be the appropriate estimator with ordinal item response data because alpha will underestimate the (theoretical underlying) reliability, especially in the binary case where the downward bias could be large. This finding is in line with those of Bandalos and Enders (1996), Jenkins and Taber

(1977), and Lissitz and Green (1975). Therefore, the ordinal coefficient alpha newly developed by Zumbo et al. (2007) is recommended for binary and ordinal data. This new statistic has been demonstrated in their study to be an accurate and stable estimator for the theoretical reliability regardless of the number of scale points and the skewed distribution of the ordinal data. However, future research needs to investigate if outliers have an effect on the estimates of this new coefficient."

At this writing, ordinal coefficients alpha and theta are not directly available in SPSS, SAS, or Stata but may be computed by them manually using the formulas given below, based on input of a polychoric correlation matrix. A polychoric correlation matrix is output by:

- The free FACTOR software developed and discussed by Lorenzo-Seva & Ferrando (2006) and available, with manual, at http://psico.fcep.urv.es/utilitats/factor/. FACTOR reads delimited ASCII data files (space, comma, or tab-delimited, with variable labels in a separate .txt file).
- SPSS has no module to calculate a polychoric correlation matrix but its factor analysis module will accept such a matrix as input. Also, Básto & Pereira (2012) have made available an SPSS R-menu program for ordinal factor analysis which produces polychoric correlation matrices.
- In SAS, the %POLYCHOR macro creates a SAS data set containing a polychoric correlation matrix or a distance matrix based on polychoric correlations.
- In Stata, the `polychoric` command estimates polychoric correlations, and the `polychoricpca` command performs principal component analysis on the resulting correlation matrix.
- Polychoric correlation matrices are also output by PRELIS, the front end to the popular structural equation modeling package LISREL, available at http://www.ssicentral.com/.
- For R, Gadermann, Guhn, & Zumbo (2012: Appendix A) describe and make available R modules for calculating ordinal coefficients alpha and theta.
- Elosua and Zumbo (2008) describe the computation or ordinal coefficient alpha using *Mplus* software.
- Other programs are listed by John Uebersax at http://www.john-uebersax.com/stat/tetra.htm#ex2 .

Ordinal coefficients alpha and theta are computed by computing a polychoric correlation matrix applying the formulas below to either the correlation coefficients or a principal components analysis respectively, using as input a matrix of polychoric correlations.

p = number of items in the scale
r_{ave} = average polychoric correlation, averaged for all pairs of items
e = the largest eigenvalue from a principal component analysis based on polychoric correlation input

Gadermann, Guhn, & Zumbo (2012: 5), provide an alternative formula for computing ordinal alpha from the correlation coefficients rather than conducting a factor analysis. This formula for ordinal coefficient alpha is:

$$\text{Ordinal coefficient alpha} = (p*r_{ave})/(1+(p-1)*r_{ave})$$

This formula is for the 'standardized alpha', an alpha that assumes that the items have equal variances. Please note, however, that unstandardized alpha and standardized alpha are the same when they are calculated from a correlation matrix such as the polychoric correlation matrix so computing the standardized alpha is an alternative method to get ordinal alpha. For a formula relating ordinal coefficient alpha to PFA, see Zumbo, Gederman, & Zeisser (2007: 22), in turn based on work by McDonald (1985: 217).

Based on Zumbo, Gadermann, & Zeisser (2007: 22), in turn based on Armor (1974: 28), the formula for ordinal coefficient theta is:

$$\text{theta} = [p/(p-1)]*[1-(1/e)]$$

Composite reliability (CR)

Also called construct reliability and Raykov's reliability rho (ρ), this coefficient tests if it may be assumed that a single common factor underlies a set of variables. Raykov (1998) has demonstrated that Cronbach's alpha may over- or under-estimate scale reliability. Underestimation is common. For this reason, rho is now preferred and may lead to higher estimates of true reliability. Raykov's reliability rho is not to be confused with Spearman's median rho, an ordinal alternative to Cronbach's alpha, discussed below. Raykov's reliability rho is output by EQS. See Raykov (1997), which lists EQS and LISREL code for computing

composite reliability. Graham (2006) discusses Amos computation of reliability rho. PLS software also computes composite reliability.

Composite reliability is a preferred alternative to Cronbach's alpha as a test of convergent validity and a measure of reliability because Cronbach's alpha may over- or under-estimate scale reliability. Underestimation is common. For this reason, composite reliability may be preferred and may lead to higher estimates of true reliability. The acceptable cutoff for composite reliability would be the same as the researcher sets for Cronbach's alpha since both attempt to measure true reliability. In an adequate model for exploratory purposes, composite reliabilities should be greater than .6 (Chin, 1998; Höck & Ringle, 2006: 15) and greater than .7 for an adequate model for confirmatory purposes. Other authors require greater than .8 (for ex., Daskalakis & Mantas, 2008: 288). Hair et al. (2010) state that in addition to CR > .7, it should also be the case that CR > AVE and AVE > 0.5 to uphold convergent validity. (See discussion of AVE above.)

Calculation of CR was discussed above in the section on discriminant validity.

Armor's reliability theta

Theta is a similar measure developed by Armor (1974). Theta =θ = [p/(p-1)]*[1-(1/λ_1)], where p = the number of items in the scale and where λ_1 denotes the first and therefore largest eigenvalue from the principal component analysis of the correlation of items comprising the scale. See Zumbo, Gadermann, & Zeisser, 2007: 22. Reliability theta is interpreted similar to other reliability coefficients. While not directly computed by SPSS or SAS, it is easily calculated from principal components factor results using the formula above.

Ordinal reliability theta. Zumbo, Gadermann, & Zeisser (2007) use a polychoric correlation matrix input to principal components analysis to calculate an ordinal version of reliability theta, using simulation studies to demonstrate ordinal reliability theta "consistently suitable estimates of the theoretical reliability, regardless of the magnitude of the theoretical reliability, the number of scale points, and the skewness of the scale point distributions." (p. 21). Ordinal reliability theta will normally be higher than the corresponding Cronbach's alpha.

Spearman's reliability rho

Spearman's rho is a form of rank order calculation. It is calculated with the same formula as for Pearson's r correlation, but using rank rather than interval data. The median rho between all pairs of items in a scale is a classic measure of reliability, in the sense of internal consistency, and as such is an ordinal alternative to Cronbach's alpha. This is not to be confused with Raykov's reliability rho.

Split-half reliability

Overview

Split-half reliability measures equivalence of instruments. It is also called parallel forms reliability or internal consistency reliability. Split-half reliability is based on administering two equivalent batteries of items measuring the same construct to the same people. If they measure the same construct they should be highly correlated. Typically these two batteries are two subsets of items from a longer instrument and are composed of items randomly selected by the computer algorithm, although it is also possible that the researcher creates the two batteries manually.

The Spearman-Brown split-half reliability coefficient is usually the measure of whether the two batteries of items are sufficiently correlated. Also called the Spearman-Brown prophecy coefficient, it is not to be confused with the Spearman-Brown stepped-up reliability coefficient (another name for standardized item alpha) discussed above. The Spearman-Brown split-half reliability coefficient is used to estimate full test reliability based on the split-half method. A common rule of thumb is that it should be .60 for exploratory purposes, .70 for adequacy for confirmatory research, and .80 or high for good reliability. Some researchers use a cutoff of .90 or higher for good reliability.

It should be noted that the Spearman-Brown coefficient is an older statistical methodology which depends upon the particular split of a larger instrument into two batteries of items. Researchers may prefer Cronbach's alpha, which does not depend on any particular split of items. Cronbach's alpha may be conceived of as the mean of all split-half reliability coefficients for all possible splits.

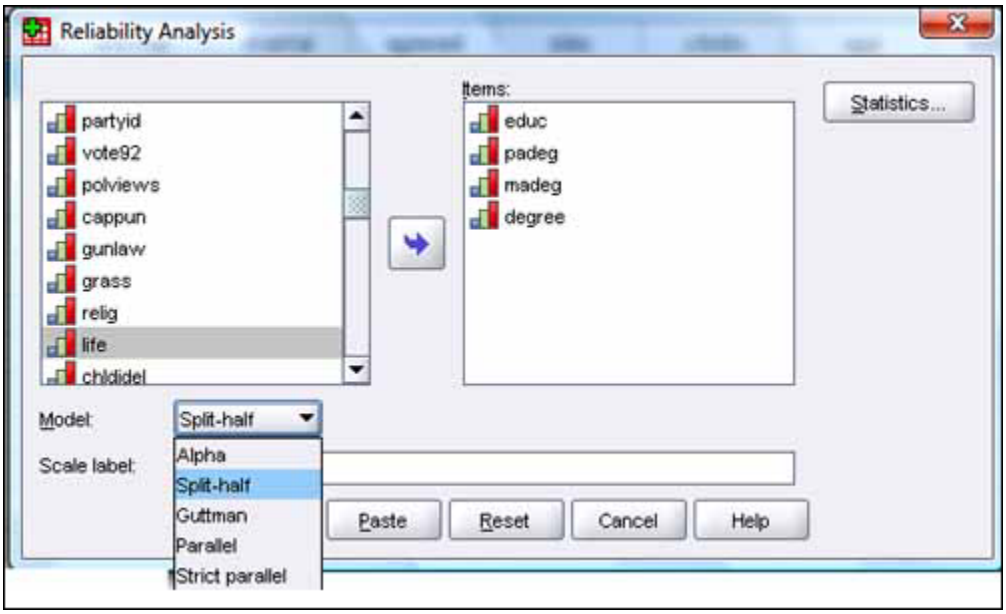
Split-half reliability in SPSS

Overview

If split half reliability is requested in SPSS in the "Model" area of the user interface illustrated above and below, four coefficients will be generated:

1. Cronbach's alpha for each form
2. The Spearman-Brown coefficient
3. The Guttman split-half coefficient
4. The Pearsonian correlation between the two forms (aka, "half-test reliability").

In SPSS, Select Analyze > Scale > Reliability Analysis; list your variables; click the "Statistics" button; select Item Scale, Scale if Item Deleted; select Split-Half from the Model drop-down list. OK. SPSS will take the first half of the items as the first split form, and the second half as listed in the dialog box as the second split form. If there are an odd number of items, the first form will be one item longer than the second form. The researcher can also use the Paste button to call up the Syntax window and alter the /MODEL=SPLIT parameter to be /MODEL=SPLIT n, where n is the number of items in the second form.



The Spearman-Brown split-half reliability coefficient

SPSS output for the Spearman-Brown coefficient is shown in the lower portion of the figure below.

Reliability Statistics

Cronbach's Alpha	Part 1	Value	.422
		N of Items	2[a]
	Part 2	Value	.534
		N of Items	2[b]
		Total N of Items	4
		Correlation Between Forms	.844
Spearman-Brown Coefficient		Equal Length	.915
		Unequal Length	.915
		Guttman Split-Half Coefficient	.794

a. The items are: Highest Year of School Completed, Father's Highest Degree.

b. The items are: Mother's Highest Degree, RS Highest Degree.

In the figure above, SPSS has divided the four-item education scale into two subscales. As shown in the table notes, the first subscale is Highest Year of School completed plus Father's Highest Degree. The second subscale is Mother's highest Degree and Respondent's Highest Degree. Comparing scores on these two subscales yields a Spearman-Brown reliability coefficient of .915. On a split-half basis, the researcher concludes the 4-item education scale is reliable.

Note that two Spearman-Brown coefficients appear in the "Reliability Statistics" table: (1) "Equal length" gives the estimate of the reliability if both halves have equal numbers of items, and (2) "Unequal length" gives the reliability estimate assuming unequal numbers of items in each battery. Here, with each battery of equal length, the coefficients are the same.

The Pearson correlation of split forms estimates the half-test reliability of an instrument or scale. The Spearman-Brown "prophecy formula" predicts what the full-test reliability would be, based on half-test correlations. This coefficient will be higher than the half-test reliability coefficient. This coefficient is usually equal to and easily hand-calculated as twice the half-test correlation divided by the quantity 1 plus the half-test reliability.

$r_{SB1} = (k^* r_{ij})/[1 + (k-1)^* r_{ij})]$
where

r_{SB1} = the Spearman-Brown split-half reliability
r_{ij} = the Pearson correlation between forms i and j
k = total sample size divided by sample size per form (k is usually 2)

As with other split-half measures, the Spearman-Brown reliability coefficient is highly influenced by alternative methods of sorting items into the two forms, which is preferably done randomly. Random assignment of items to the two forms should assure equality of variances between the forms but this is not guaranteed and should be checked by the researcher.

The Guttman split-half reliability coefficient and Guttman's lower bounds (lambda 1 – 6)

The Guttman split-half reliability coefficient is an adaptation of the Spearman-Brown coefficient, but one which does not require equal variances between the two split forms. Select Scale in the Descriptives area of the Statistics button to generate the Scale Statistics table, notes to which will list which items are in each of the two subscales created automatically by the SPSS split half algorithm, similar to that in the "Reliability Statistics" table illustrated above for Spearman-Brown split-half analysis. Guttman's split-half reliability is lambda 4 in the figure below.

Guttman advocated use of the Guttman split-half reliability in conjunction with five other coefficients. Guttman's lower bounds refers to a set of six lambda coefficients, L1 to L6, generated when in SPSS one selects "Guttman" under the Model button. Below is output for the four-item education scale discussed previously.

Reliability Statistics

Lambda	1	.531
	2	.753
	3	.707
	4	.794
	5	.795
	6	.856
N of Items		4

L1: An intermediate coefficient used in computing the other lambdas.

L2: More complex than Cronbach's alpha and preferred by some researchers, though less common.

L3: Equivalent to Cronbach's alpha.

L4: Guttman split-half reliability.

L5: Recommended when a single item highly covaries with other items, which themselves lack high covariances with each other.

L6: Recommended when inter-item correlations are low in relation to squared multiple correlations

Guttman recommended experimenting to find the split of items which maximizes Guttman split-half reliability (L4), then for that split using the highest of the lower bound lambdas as the reliability estimate for the set of items.

Split-half reliability in SAS

SAS does not support direct computation of either Spearman-Brown or Guttman split-half reliability. SAS program for Split-Half reliability. A third-party company, Psychometric & Statistical Solutions, sells two SAS macros at its website, http://psychometricmethods.com/reliability.html:

1. SAS program for split-half reliability: This program uses SAS arrays for simultaneously deriving split-half reliability for 12 tests and 7 age groups.
2. SAS significant difference for reliabilities: This program uses SAS macros to compute significant differences for large numbers of tests at a specified alpha level and produces an output table for publication.

Split-half reliability in Stata

Stata does not support either Spearman-Brown or Guttman split-half reliability. However, as discussed above, a third party has made available the sbrowni.ado file, which performs a Spearman-Brown reliability correction for test length but does not compute the Spearman-Brown coefficient . Likewise, the third-party program sbri, discussed above, computes a type of Spearman-Brown coefficient

used in conjunction with regression models but not of the ordinary split-half reliability type.

To this author's knowledge, there is no program supporting Guttman split-half reliability in Stata.

Odd-Even Reliability

Overview

Odd-even reliability is a variant of split half reliability where the two test item batteries are created by the researcher from the odd-numbered and even-numbered items respectively.

Odd-even reliability in SPSS

SPSS supports odd-even reliability in syntax but not in the menu system. For the four-item education scale, the split-half syntax code (accessed by clicking the Paste button) was:

```
DATASET ACTIVATE DataSet1.
RELIABILITY
  /VARIABLES=educ padeg madeg degree
  /SCALE('ALL VARIABLES') ALL
  /MODEL=SPLIT
  /STATISTICS=SCALE ANOVA.
```

By default, the SPSS split half algorithm makes the first two items (educ, padeg) into subscale 1 and makes the last two items (madeg, degree) into subscale 2. To convert to odd-even split half format, the syntax of the /VARIABLES statement must be changed as illustrated below:

```
/VARIABLES=educ madeg padeg degree
```

Note that the order of the variables has been rearranged to list the 1st and 3rd variables first, followed by the 2nd and 4th. This will make subscale 1 be the odd items and subscale 2 be the even items. As a result, the Spearman Brown split half reliability coefficient will differ but the ANOVA table for the model as a whole will not change.

Reliability Statistics

Cronbach's Alpha	Part 1	Value	.380
		N of Items	2[a]
	Part 2	Value	.537
		N of Items	2[b]
		Total N of Items	4
		Correlation Between Forms	.805
Spearman-Brown Coefficient		Equal Length	.892
		Unequal Length	.892
		Guttman Split-Half Coefficient	.814

a. The items are: Highest Year of School Completed, Mother's Highest Degree.

b. The items are: Father's Highest Degree, RS Highest Degree.

Odd-even reliability in SAS and Stata

SAS and Stata do not support odd-even reliability, though, of course, the researcher could create odd and even item subscales manually and the use the correlation procedure on them.

Test-retest reliability

Test-retest reliability, which measures stability over time, is administering the same test to the same subjects at two points in time, then correlating the results. The appropriate length of the interval depends on the stability of the variables which causally determine that which is measured. A year might be too long for an opinion item but be appropriate for a physiological measure. A typical interval is several weeks. Statistically, test-retest reliability is treated as a variant of split-half reliability and also uses the Spearman-Brown coefficient.

Test-retest methods are disparaged by many researchers as a way of gauging reliability. Among the problems are that short intervals between administrations of the instrument will tend to yield estimates of reliability which are too high. There may be invalidity due to a learning/practice effect (subjects learn from the first administration and adjust their answers on the second). There may be invalidity due to a maturation effect when the interval between administrations is long (the subjects change over time). The bother of having to take a second administration may cause some subjects to drop out of the pool, leading to

nonresponse biases. Note, however, that test-retest designs are still widely used and published and there is support for this. McKelvie (1992), for instance, reports that reliability estimates under test-retest designs are not inflated due to memory effects. Researchers using test-retest reliability must address the special validity concerns but may decide to go proceed if warranted.

Inter-rater reliability

Overview

Inter-rater reliability, which measures homogeneity across raters who are performing measurements, is administering the same form to the same people by two or more raters/interviewers so as to establish the extent of consensus on use of the instrument by those who administer it. Raters should be as blinded as possible to expected outcomes of the study and should be randomly assigned. In the data setup for most statistical packages, judges are the columns and subjects are the rows. For categorical data, consensus is measured as number of agreements divided by total number of observations. For continuous data, consensus is measured by intraclass correlation, discussed below.

Cohen's kappa

Overview

Cohen's kappa can be used to assess inter-rater reliability for two raters. Cohen developed a version of kappa for more raters but it is not implemented as such in SPSS. However, kappa weighted for multiple raters is equivalent to the default form of intraclass correlation, ICC (see below).

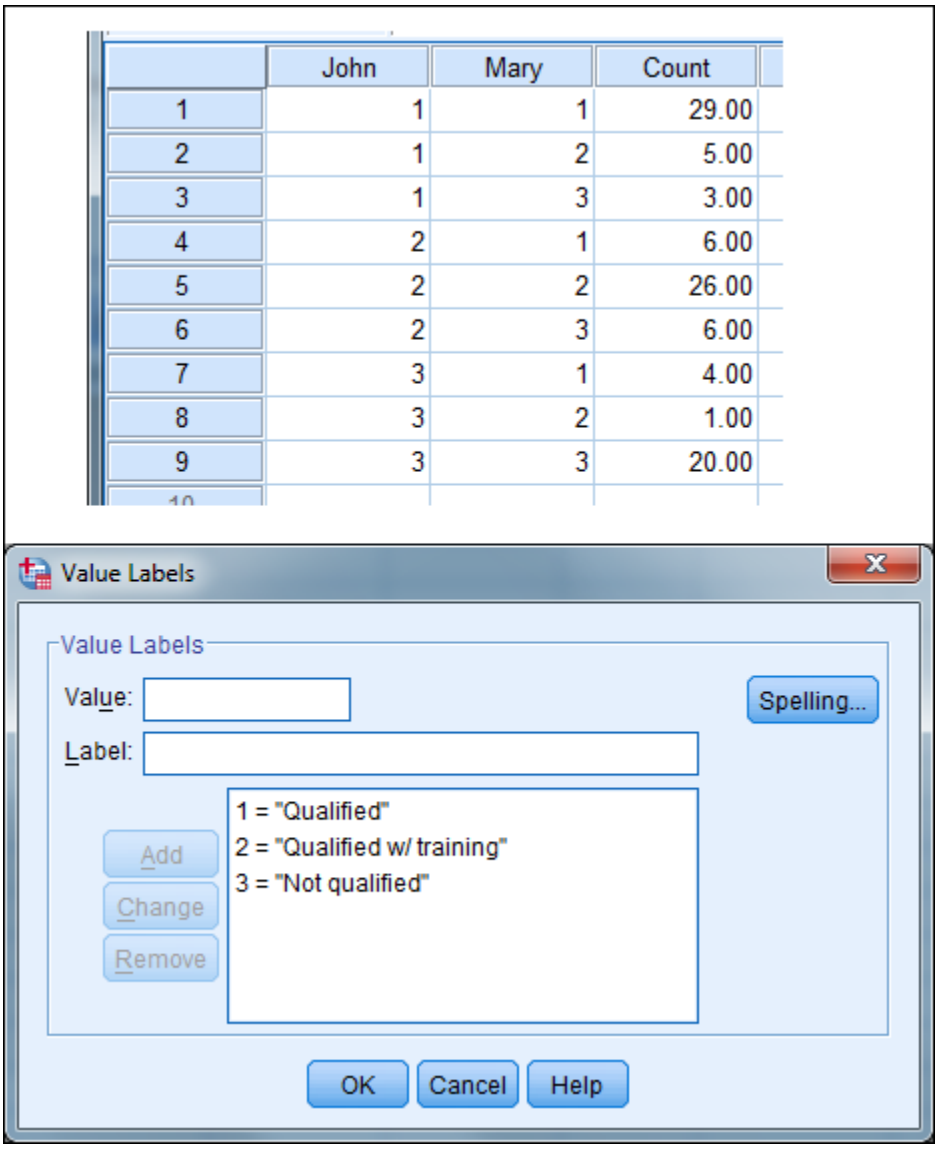
Kappa assumes the objects of ratings are independent (the rating of one does not affect the rating of another), that the raters are independent (one rater's ratings do not affect those of another rater), and that all ratings are by the same two raters. The ratings may be categorical or binned continuous data.

Kappa in SPSS

Kappa is available in the Crosstabs procedure in SPSS: select Analyze > Descriptive Statistics >Crosstabs, then click the Statistics button to select kappa (it is not a default option).

Example

John and Mary rate 100 job applications as 1 = Qualified, 2 = Qualified with training, or 3 = Unqualified. The data setup and coding is shown in the figure below, where the "Value Labels" portion shows the coding. Count is a weight variable corresponding to the table further below.



Crosstabs output is shown below, with kappa reported in the "Value" column of the "Symmetric Measures" table.

John * Mary Crosstabulation

Count

		Mary			
		Qualified	Qualified w/ training	Not qualified	Total
John	Qualified	29	5	3	37
	Qualified w/ training	6	26	6	38
	Not qualified	4	1	20	25
Total		39	32	29	100

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.622	.065	8.796	.000
N of Valid Cases		100			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Interpretation

By convention, kappa > .70 is considered acceptable inter-rater reliability, but this depends highly on the researcher's purpose. Some accept > .60, particularly for exploratory purposes. Another rule of thumb is that K = 0.40 to 0.59 is moderate inter-rater reliability, 0.60 to 0.79 substantial, and 0.80 outstanding (Landis & Koch, 1977). For inter-rater reliability of a set of items, such as a scale, one would report mean Kappa.

For the examples above, at Kappa = .622, inter-rater reliability between John and Mary is adequate for exploratory purposes but lacks substantial mutual agreement. Note that while significance levels are reported, these may be interpreted only if applicants represent a random sample.

Algorithm. For a table of Rater A vs. Rater B, let a = the sum of counts on the diagonal, reflecting agreements. Let e = the sum of expected counts on the

diagonal, where expected is calculated as [(row total * column total)/n], summed for each cell on the diagonal. Let n = the total number of ratings (observations). Kappa then equals the ratio of the surplus of agreements over expected agreements, divided by the number of expected disagreements. This is equivalent to K = (a - e)/(n - e). Fleiss and Cohen (1973) have shown ICC, discussed below, is mathematically equivalent to weighted kappa. For ordinal rankings or better, one may weight each cell in the agreement/disagreement table by a weight between 0 and 1, where 1 corresponds to the row and column categories being the same and 0 corresponds to the categories being maximally dissimilar.

Kappa in SAS

In SAS, both simple and weighted kappa are computed by the FREQ procedure.

SAS syntax

The LIBNAME statement indicates the directory where the data file is stored. The directory is given the working name "kappa".

```
LIBNAME kappa "C:\Data";
```

PROC CONTENTS is an optional statement which prints out basic information about the data, not shown below. It is useful for verifying that the data are as intended.

```
PROC CONTENTS DATA=kappa.reliab_kappa;
RUN;
```

The PROC FREQ statement launches SAS's crosstabulation program. The TABLES statement asks for a table of the two variables, John and Mary, which contain their respective ratings. The "/ agree" option requests tests and measures of classification agreement, discussed in output below. (There is a long list of other output options, beyond out scope here). The TEST statement requests default output for Cohen's kappa, discussed below. Because the data were input in weighted form, as illustrated above for SPSS, the WEIGHT statement is used to invoke weighting by the variable Count, which contains cell frequencies.

```
PROC FREQ DATA = kappa.reliab_kappa;
TABLES John * Mary / agree;
TEST kappa;
WEIGHT Count;
```



```
RUN;
```

SAS output: The crosstabulation table

Below is the default crosstabulation table, which merely reproduces the table of agreements and disagreements. The diagonal contains agreements between the two raters and contains most of the ratings. All off-diagonal cells represent disagreements.

The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of John by Mary			
	Mary			
John	1	2	3	Total
1	29	5	3	37
	29.00	5.00	3.00	37.00
	78.38	13.51	8.11	
	74.36	15.63	10.34	
2	6	26	6	38
	6.00	26.00	6.00	38.00
	15.79	68.42	15.79	
	15.38	81.25	20.69	
3	4	1	20	25
	4.00	1.00	20.00	25.00
	16.00	4.00	80.00	
	10.26	3.13	68.97	
Total	39	32	29	100
	39.00	32.00	29.00	100.00

SAS output: The test of symmetry table

The table below gives the S statistic, which tests for symmetry in the table, and gives the corresponding probability level. This is Bowker's test for symmetry, applicable to tables where there are more than two categories (here there were

three). Had there been only two rating categories, McNemar's test would have been reported, interpreted similarly.

That the symmetry test is non-significant below means that the researcher fails to reject the null hypothesis that cell proportions in the table are symmetric. Conversely, if S is significant, the researcher concludes the marginals are not homogenous. Homogenous marginals indicate that the two raters share the same propensity across ratings categories. However, even if S is significant and the raters appear to differ, this does not invalidate kappa as a measure of inter-rater reliability.

Statistics for Table of John by Mary

Test of Symmetry	
Statistic (S)	3.8052
DF	3
Pr > S	0.2833

SAS output: The kappa tables

As illustrated below, the AGREE option also produces the simple kappa coefficient and the weighted kappa coefficient, with standard errors and confidence limits (in the tables below, ASE is asymptotic standard error). When there are only two rating categories, only simple kappa is reported since weighted kappa is the same.

For simple kappa a significance test is reported. The p value here is significant, meaning that kappa differs significantly from 0.

If there are more than two categories, as in this example, the weighted kappa coefficient is also reported. Weighted kappa is a generalization of the simple kappa coefficient which uses weights to quantify the relative difference between categories. It is different from simple kappa, which is what SPSS reported for the same data in the example above. Weighted kappa is used in preference to simple kappa if the researcher feels that close differences (ex., John rates a candidate 1, Mary rates the candidate 2) should be weighted more than greater differences (ex., John rates 1, Mary rates 3). Weighted kappa may be higher than simple kappa, as for these data, thus reporting higher reliability. The significance of weighted kappa is the same as for simple kappa, hence is not reported.

When there are multiple strata (not the case in this example), the AGREE option also provides tests for equal kappas among strata. If there are multiple strata and only two response categories, PROC FREQ computes Cochran's test. If there are multiple strata, PROC FREQ combines the stratum-level estimates of kappa into an overall estimate of a common value of kappa.

Simple Kappa Coefficient	
Kappa	0.6221
ASE	0.0653
95% Lower Conf Limit	0.4942
95% Upper Conf Limit	0.7500

Test of H0: Kappa = 0	
ASE under H0	0.0707
Z	8.7958
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Weighted Kappa Coefficient	
Weighted Kappa	0.6307
ASE	0.0690
95% Lower Conf Limit	0.4954
95% Upper Conf Limit	0.7659

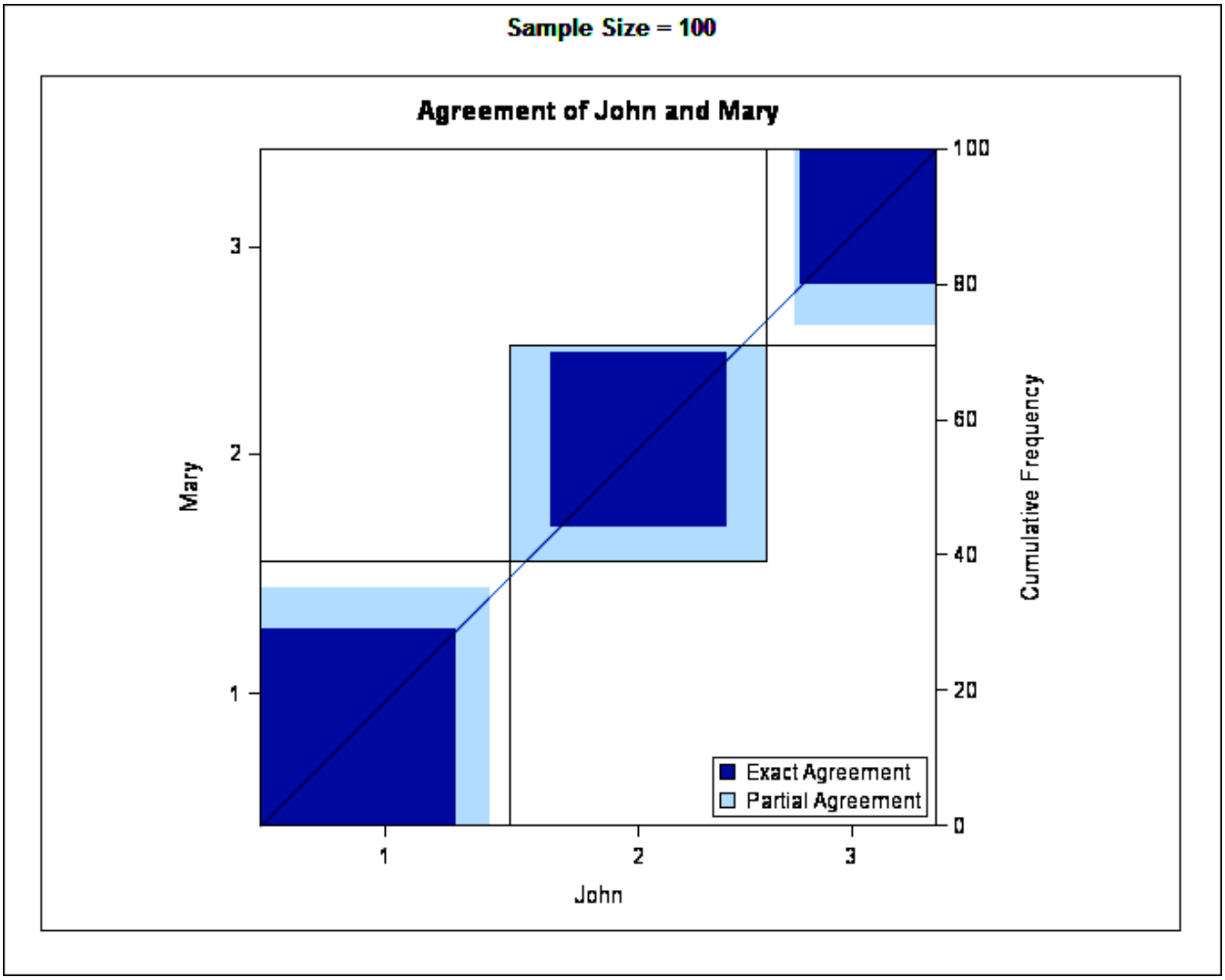
Sample Size = 100

SAS output: The agreement plot

The agreement plot for the example data is illustrated in the figure below (contrast has been increased to differentiate white/light blue/dark blue). "Exact agreement" in the plot corresponds to the diagonal cells of the square crosstabulation shown above and is shown by the dark blue squares in the plot. "Partial agreement" corresponds to the adjacent off-diagonal table cells, where

the row and column values are within one level of exact agreement, and is shown by the light blue areas. Disagreements greater than one level are shown in white.

Thus in each segment, the larger the dark blue area, the more the two raters are agreeing. The larger the light blue area, the more the two raters are partially agreeing. The larger the white area, the more the two raters are disagreeing by more than one level in the rating coding scheme.



Kappa in Stata

Overview

In Stata, the kappa command computes Cohen's kappa for interrater agreement. However, this command has four flavors, the syntax for which is given below:

1. kap varname1 varname2 [if] [in] [weight] [, options]
    a. Interrater agreement, two unique raters. This flavor is the one wanted here to match SPSS results discussed above.
2. kap varname1 varname2 varname3 [...] [if] [in] [weight]
    a. Interrater agreement, nonunique raters, variables record ratings for each rater
3. kappa varlist [if] [in]
    a. For interrater agreement, nonunique raters, variables record frequency of ratings. Options are not allowed.
4. kapwgt wgtid [1 \ # 1 [\ # # 1 ...]]
    a. Weights for weighting disagreements

Unfortunately, the kap and kappa commands are among the few Stata commands which do not support weighting of cases so for the example below, an expanded version of the data file described above for SPSS was used, without need for the Count variable.

Stata syntax and output

The command "kap John Mary, tab" produces the output shown below. The "tab" option produces the crosstabulation, which Stata calls the "table of assessments". The resulting Kappa or .622 is the same as in SPSS, with the same interpretation.

```
. kap John Mary, tab
```

John	Mary Qualified	Qualified	Not quali	Total
Qualified	29	5	3	37
Qualified w/ training	6	26	6	38
Not qualified	4	1	20	25
Total	39	32	29	100

Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
75.00%	33.84%	0.6221	0.0707	8.80	0.0000

Kendall's coefficient of concordance, W

Overview

Kendall's coefficient of concordance, W, may be used as a measure of inter-rater reliability for multiple raters. It is closely related to Friedman's test but is preferred because Kendall's W is normed from 0 to 1, with 0 meaning no agreement across raters and 1 meaning perfect agreement. Agreement, of course, corresponds to correlation of the ratings for rated objects across raters. The same cutoffs apply as for other measures of reliability: .6 for exploratory research, .7 adequate for confirmatory research , .8 good inter-rater reliability for confirmatory research. Some researchers require the more stringent cutoff of .9 for confirmatory research.

Kendall's coefficient of concordance was designed for the ordinal level of measurement. It may be used for interval data but will have less power than measures which assume interval data (ex., intraclass correlation).

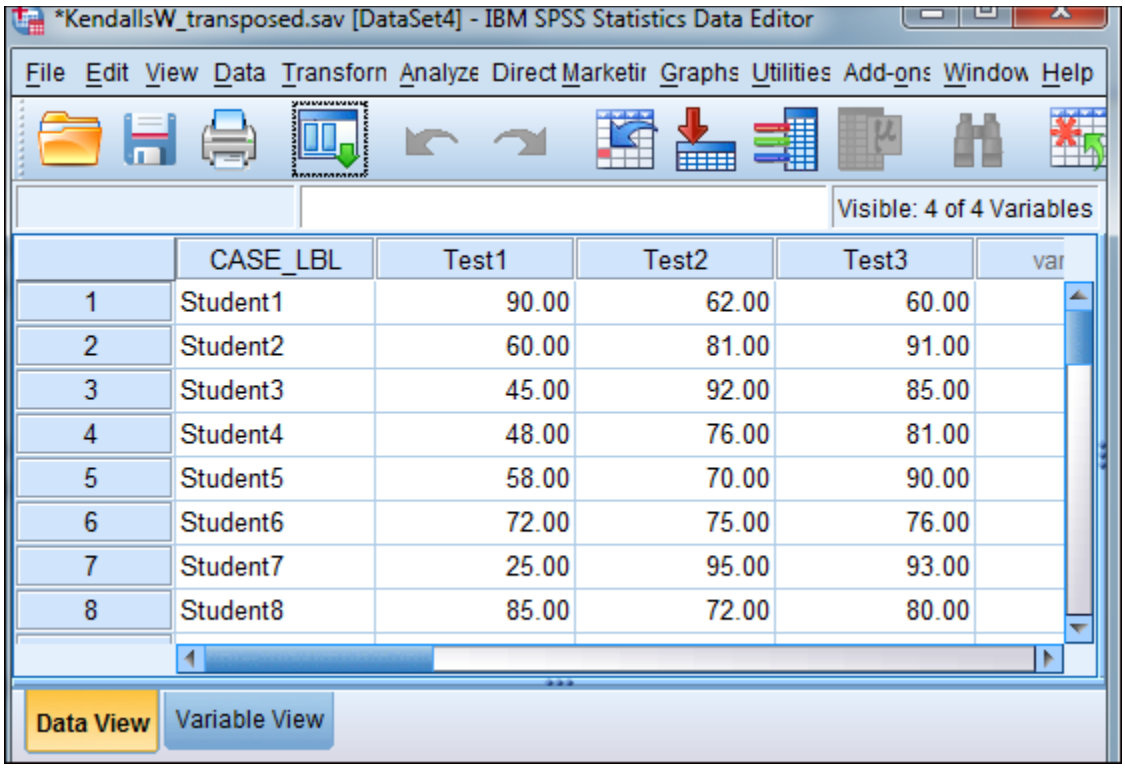
If W is significant, the researcher concludes that there is significant inter-rater correlation. If W is non-significant, the researcher fails to reject the null hypothesis is that inter-rater correlation is zero (that there is no agreement across raters). Like all significance tests, non-significance may be due to a very small sample as well as to insufficient inter-rater agreement.

Put another way, if W is significant the researcher concludes that the distributions of raters on the ratings are drawn from the different underlying populations. If W is non-significant, the researcher fails to reject the null hypothesis that the raters' ratings are drawn from the same underlying population.

Example data setup

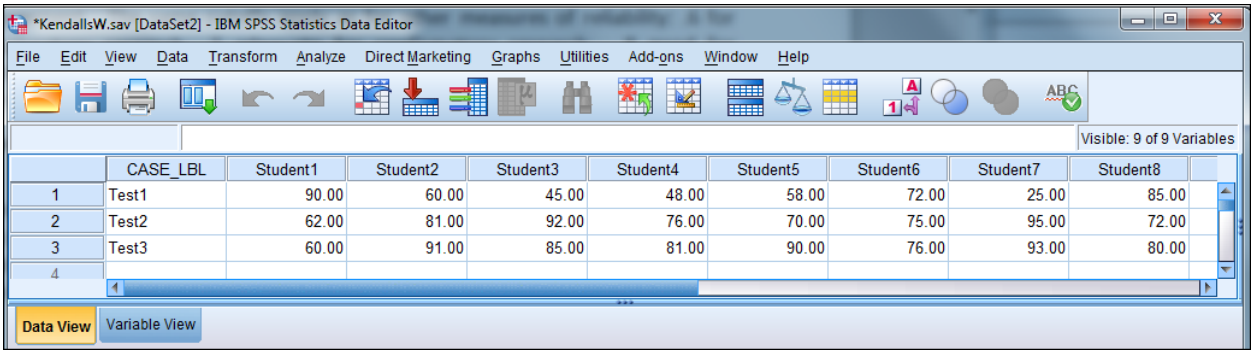
The example below uses data supplied with the Kendall W's program file in Stata. It can be typed in easily by the reader as it consists of just eight students being rated on three tests. Kendall's W assesses the extent to which students receive similar scores on each of the three tests. Data might be in one of two forms.

Raters are rows. In this format, raters are rows and objects are columns. For the example, students are rows and tests are columns. Cell entries are the test scores. SPSS wants this format. This format must be transposed first to use in Stata.



	CASE_LBL	Test1	Test2	Test3	var
1	Student1	90.00	62.00	60.00	
2	Student2	60.00	81.00	91.00	
3	Student3	45.00	92.00	85.00	
4	Student4	48.00	76.00	81.00	
5	Student5	58.00	70.00	90.00	
6	Student6	72.00	75.00	76.00	
7	Student7	25.00	95.00	93.00	
8	Student8	85.00	72.00	80.00	

Raters are columns. In this format raters are columns and objects are rows. For the example, tests are columns and students are rows. Cell entries are the test scores. Stata wants this format. This format must be transposed first to use in SPSS.



	CASE_LBL	Student1	Student2	Student3	Student4	Student5	Student6	Student7	Student8
1	Test1	90.00	60.00	45.00	48.00	58.00	72.00	25.00	85.00
2	Test2	62.00	81.00	92.00	76.00	70.00	75.00	95.00	72.00
3	Test3	60.00	91.00	85.00	81.00	90.00	76.00	93.00	80.00
4									

Transposition. In SPSS, a dataset is transposed data by selecting Data > Transpose. In Stata, use the xpose command.

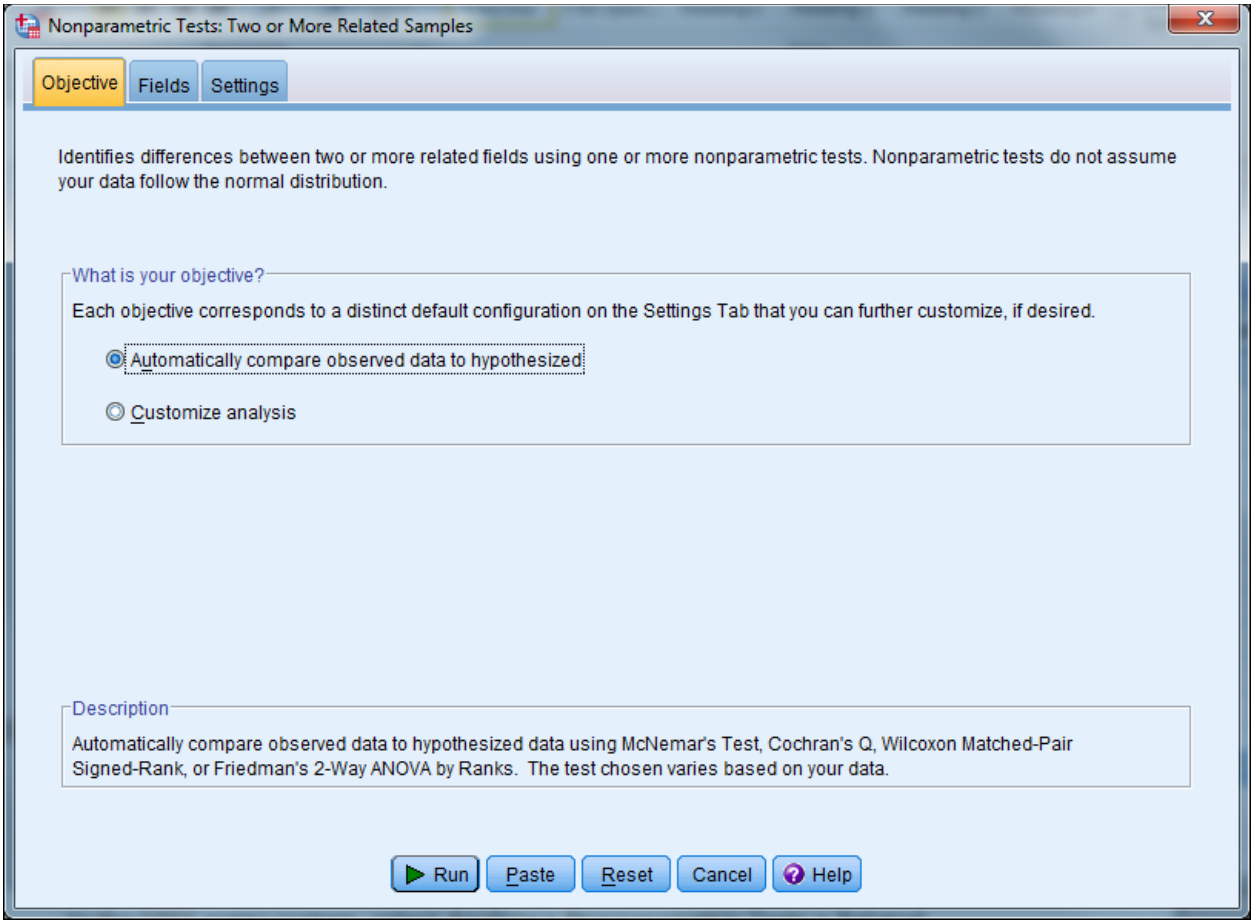
Kendall's W in SPSS

SPSS setup

In the SPSS menu system, select Analyze > Nonparametric Tests > Related Samples. A three-part dialog comes up.

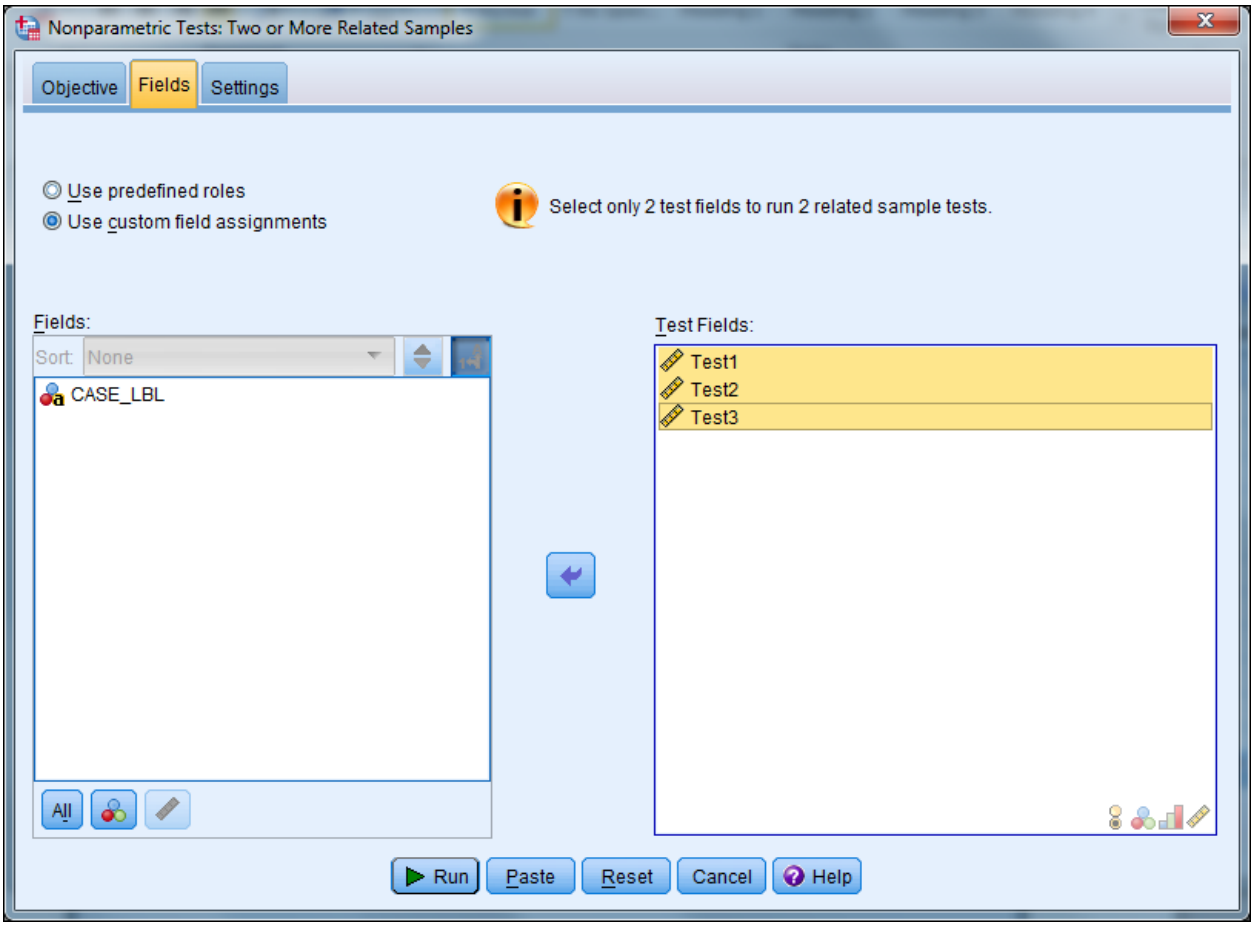
Tab 1: Objectives

The default may be accepted for this tab.



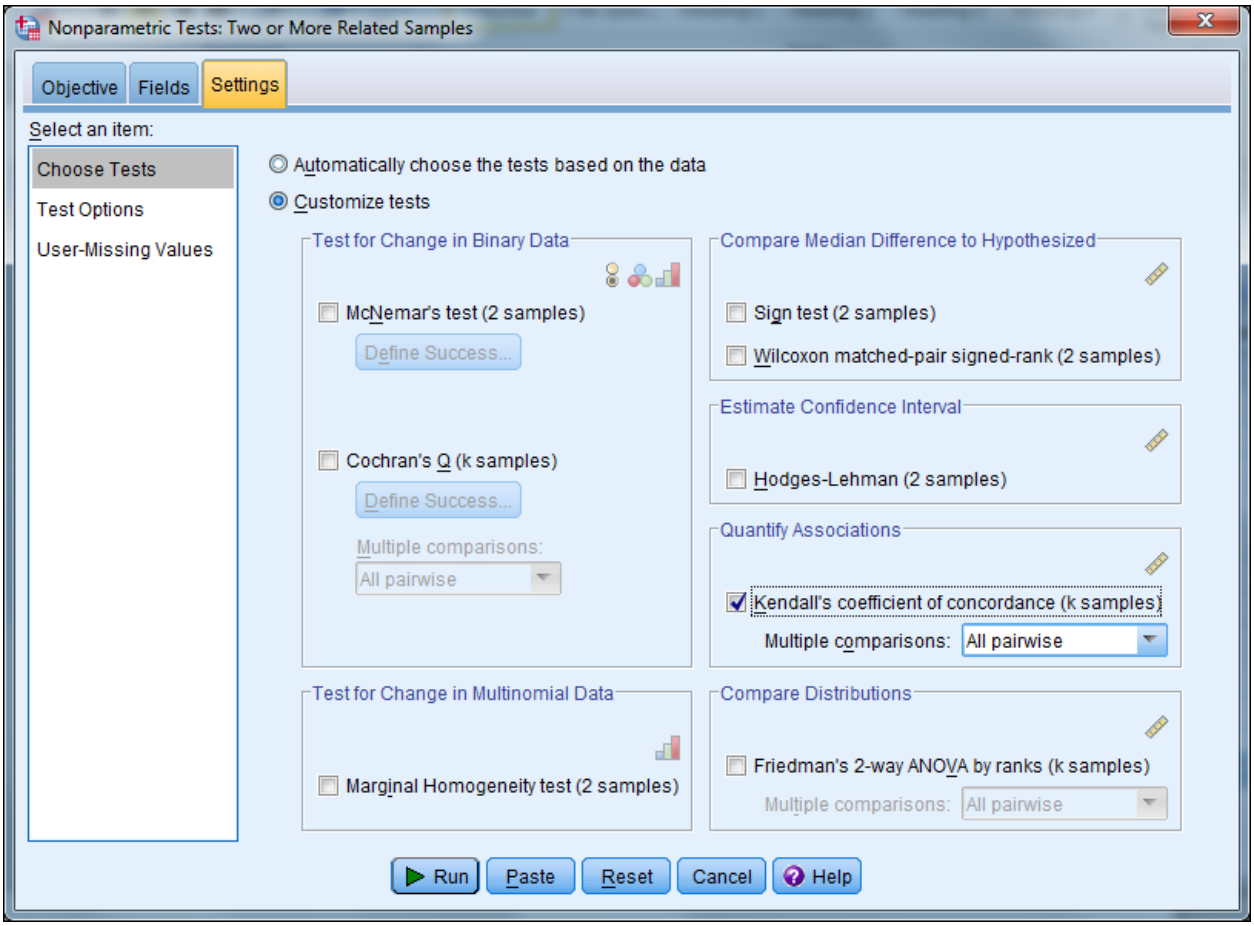
Tab 2: Fields

Move the objects (here, Tests) over into the "Test Fields" area.



Tab 3: Settings

This tab has three items. In the "Choose Tests" item, click "Customize tests" and select "Kendall's coefficient of concordance." The "Test Options" item, not shown, would allow the researcher to set an alpha significance cutoff of other than the default .05 and confidence limits of other than the default .95. The "User-Missing Values" item, not shown, allows the default of excluding user-missing values for categorical variables from analysis to instead include them. User-missing values for continues variables are always excluded.



SPSS output

Basic SPSS output shows the summary table below. For the example data the "Hypothesis Test Summary" table shows that Kendall's W is non-significant. That W is non-significant means that the researcher fails to reject the null hypothesis is that inter-rater correlation is zero (that there is no agreement across raters). Below, it is shown that W is only .203. Given the very small sample size (8), it cannot be concluded that .203 is different from 0. SPSS prints out the equivalent but perhaps more obscure statement that "The distributions of Test1, Test2, and Test3 are the same."

Nonparametric Tests

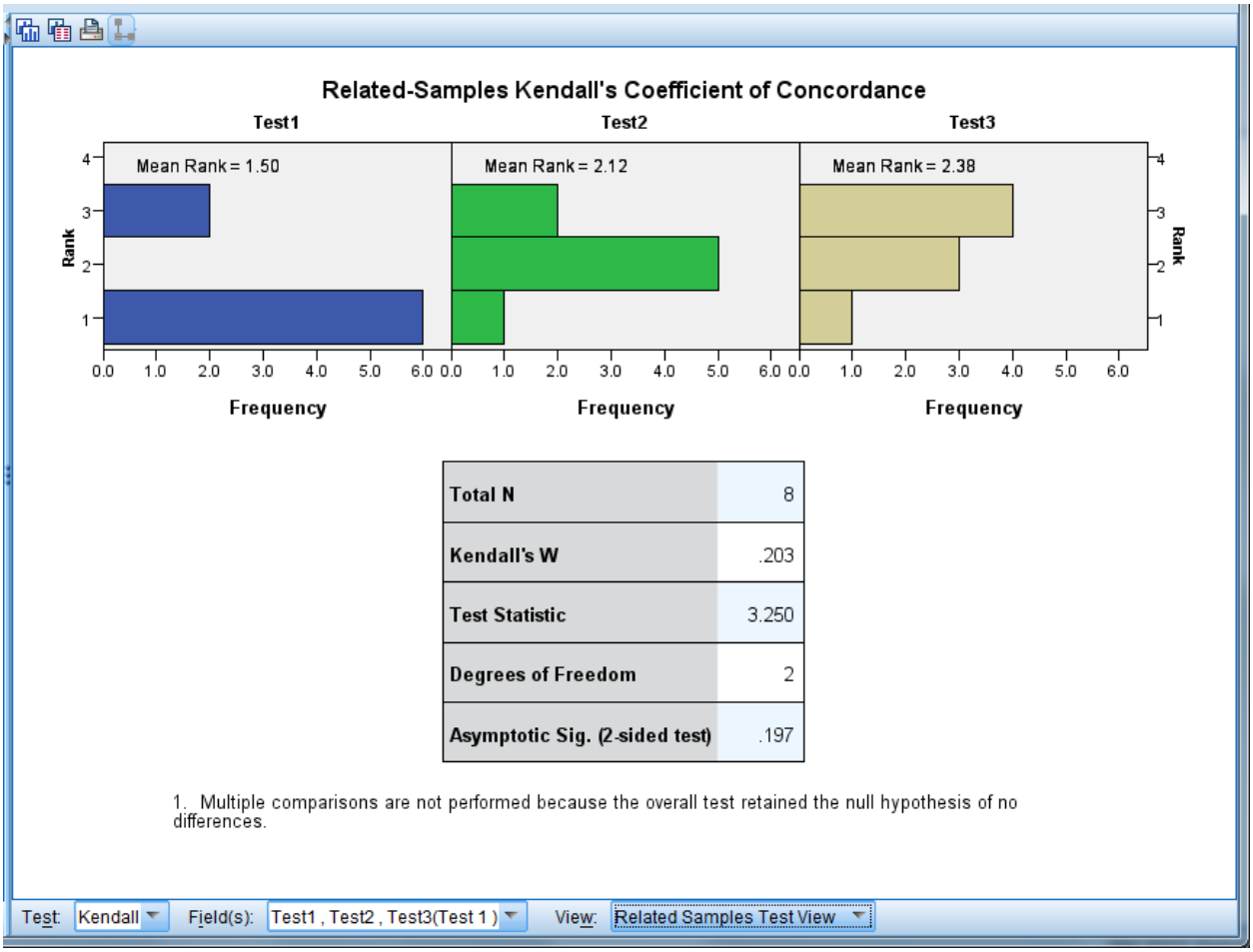
[DataSet10] C:\Data\KendallsW_Raters_Are_Rows.sav

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Test1, Test2 and Test3 are the same.	Related-Samples Kendall's Coefficient of Concordance	.197	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Double-clicking on the table above in output brings up the SPSS "Model Viewer," with more detailed output shown further below.



In the Model Viewer table above, Kendall's W is relatively low (.203) , reflecting relative low inter-rater agreement, which here is inter-student similarity on test scores. Given low sample size (8), it is so low it cannot be said to be significantly different from 0 (p = .197).

Kendall's W in SAS

SAS does not directly support Kendall's coefficient of concordance (ex., W is not a test in PROC FREQ, though kappa is). However, third parties have written SAS macros for Kendall's W, including the MAGREE macro (Go to http://support.sas.com and search for "magree macro").

Kendall's W in Stata

Stata setup

In Stata, Kendall's W is implemented through a third-party .ado file by Richard Goldstein, based on work by Gibbons (1985). The program can be installed with the command "**`net install snp2_1`**". This will install the friedman.ado module, after which one may type "help kendall" to view information about the coefficient of concordance aspect of the program.

After installing friedman.ado and putting the data file described above in use, Kendall's W can be calculated by issuing the command below. Note that Stata wants the data to be in raters-are-rows format, otherwise it must be transposed, which is done by way of illustration below. Note the use of the "?" as a wildcard operator expecting that the objects being rated will all have the same root name, here "Test".

Stata output

If raters-are-rows data are used, the `xpose` command must be used to transpose the data as shown in the upper portion of the figure below. If raters-are-columns data are used, the `friedman` command may be issued directly, as shown in the lower portion of the figure below.

Kendall's coefficient of concordance and its significance level are the same as in SPSS, with the same interpretation given above for SPSS output.

```
. use "C:\Docs\David\Statistical Associates\Data\KendallsW_Raters_Are_Rows.dta", clear

. xpose, clear

. friedman v?

Friedman =    3.2500
Kendall  =    0.2031
P-value  =    0.1969


. use "C:\Docs\David\Statistical Associates\Data\KendallsW_Raters_Are_Columns.dta",clear

. friedman Student?

Friedman =    3.2500
Kendall  =    0.2031
P-value  =    0.1969
```

We start with raters-are-rows format.
The xpose command transposes the data with row labels becoming v1, v2, v3, etc.

We obtain Kendall's coefficient of concordance with the friedman command, referencing the new v? variables, which are the eight students.

If we entered raters-are-columns formatted data, with columns labeled Student1, Student2, Student3, etc., and with the rows being the three tests, we do not need to transpose.

That Kendall's W is low means that there is only a weak tendency for students to score the same on the three tests. That its probability value is non-significant means that the researcher fails to reject the null hypothesis that inter-rater agreement among students is 0, though this finding in part reflects the very low sample size (8).

Intraclass correlation (ICC)

Overview

Intraclass correlation (ICC) is used to measure inter-rater reliability for two or more raters when data may be considered interval level. It may also be used to assess test-retest reliability. ICC may be conceptualized as the ratio of between-groups variance to total variance, as elaborated below. A classic citation for intraclass correlation is Shrout and Fleiss (1979), though ICC is based on work going back before WWI.

ICC is sometimes used outside the context of inter-rater reliability. In general, ICC is a coefficient which approaches 1.0 as the between-groups effect (the row effect) is very large relative to the within-groups effect (the column effect), whatever the rows and columns represent. In this way ICC is a measure of homogeneity: it approaches 1.0 when any given row tends to have the same values for all columns. For instance, let columns be survey respondents and let rows be Census block numbers, and let the attribute measured be

white=0/nonwhite=1. If blocks are homogenous by race, any given row will tend to have mostly 0's or mostly 1's, and ICC will be high and positive. As a rule of thumb, when the row variable is some grouping or clustering variable, such as Census areas, ICC will more and more approach 1.0 as the size of the clusters decreases and becomes more compact (ex., as one goes from metropolitan statistical areas to Census tracts to Census blocks). ICC is 0 when within-groups variance equals between-groups variance, indicative of the grouping variable having no effect. Though less common, note that ICC can become negative when the within-groups variance exceeds the between-groups variance.

Interpretation

ICC is interpreted similar to kappa, discussed above. ICC will approach 1.0 when there is no variance within targets (for any target, all raters give the same ratings), indicating total variation in measurements is due solely to the target (ex., TV attribute) variable. That is, ICC will be high when any given row tends to have the same score across the columns (which are the raters). For instance, one may find all raters rate an item the same way for a given target, indicating total variation in the measure of a variable depends solely on the values of the variable being measured -- that is, there is perfect inter-rater reliability. Put another way, ICC may be thought of as the ratio of variance explained by the independent variable divided by total variance, where total variance is the explained variance plus variance due to the raters plus residual variance. ICC is 1.0 only when there is no variance due to the raters and no residual variance to explain.

ICC in multilevel models

Intraclass correlation of a different nature is implemented in multilevel models (a.k.a. linear mixed models or hierarchical linear models). “Multilevel” might refer, for instance, to students nested within classrooms. Multilevel modeling will predict some student attribute (ex., math scores) based on student-level variables and as adjusted for a classroom effect (and possibly other level 2 variables). ICC in this context is not inter-rater reliability. Rather ICC is a measure of how large a proportion between-group (between-classroom) variance is as a percent of the sum of between-group variance plus within-group (residual) variance.

```
ICC = between-group variance/(between-group variance + within-group variance)
```

In SPSS, this type of ICC may be calculated under the menu choices Analyze > Mixed Models > Linear or Analyze > Mixed Models > Generalized Linear.

In SAS, PROC NLMIXED will compute the intraclass correlation coefficient (ICC) for a continuous of binary dependent variable and PROC MIXED will compute ICC for continuous dependent variables.

In Stata, ICC is generated by the "estat icc" postestimation command following the xtmixed command, which implements linear mixed modeling.

This type of ICC is for assessing intragroup similarity with respect to a single dependent variable and is to be distinguished from ICC as a measure of inter-rater agreement as discussed here.

Sample size: ICC vs. Pearson r

When there are just two ratings, ICC is preferred over Pearson's r only when sample size is small (< 15). Since Pearson's r makes no assumptions about rater means, a t-test of the significance of r reveals if inter-rater means differ. For small samples (< 15), Pearson's r overestimates test-retest correlation and in this situation, intraclass correlation is used instead of Pearson's r. Walter, Eliasziw, & Donner (1998) set optimal sample size for ICC based on desired power level, magnitude of the predicted ICC, and the lower confidence limit, concluding that if the researcher used the customary .95 confidence level and the .80 power level, and had two ratings per subject, then the needed sample size (needed to prove the estimated ICC was different from 0) would range from 5 when the estimated ICC was .9 to 616 when it was only .1; for three ratings, the corresponding range was 3 to 225; for four ratings, 3 to 123; for five ratings, 3 to 81; for 10 ratings, 2 to 26; for 20 ratings, 2 to 11 (pp. 106-107).

Bonnett (2002: 1334) investigated the sample size issue for ICC, concluding that optimum sample size is a function of the size of the intraclass correlation coefficient and the number of ratings per subject, as well as the desired significance level (alpha) and desired width (w) of the confidence interval. For alpha = .95 and w=.2, Bonnett concluded that the optimal sample size for two ratings varied from 15 for ICC=.9 to 378 for ICC = .1; for three ratings, it varied from 13 to 159; five ratings, 10 to 64; and 10 ratings, 8 to 29. That is, the fewer ratings and the smaller the ICC level, the larger the needed sample size. For the

example used below, with 906 people rating 7 items, described above, sample size is more than adequate.

Example data

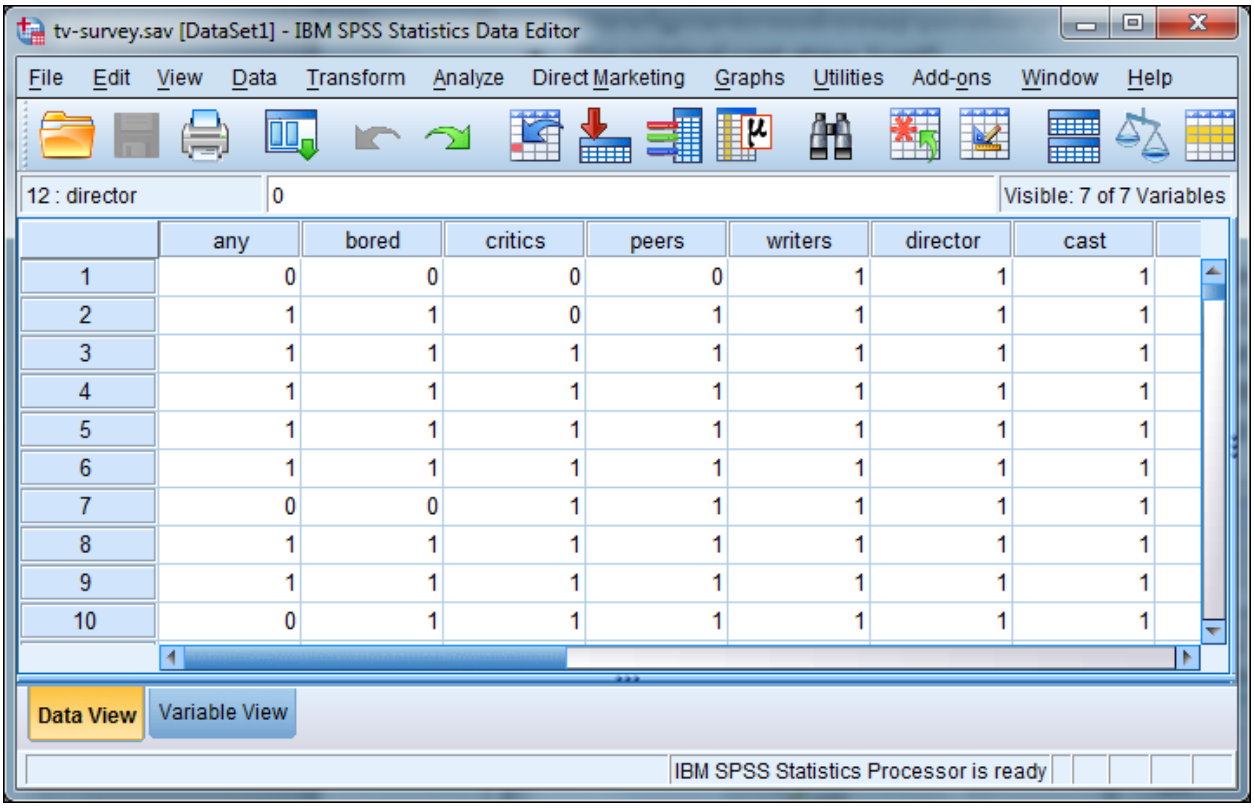
The example below uses the data file tv-survey.sav, supplied with SPSS as a sample. Conversion to SAS or Stata format was discussed above. Some 906 respondents were asked if they would watch a particular show for any of the reasons shown below (variable names shown in parentheses). Items were non-exclusive (raters responded affirmatively (coded 1) or negatively (coded 0) to each item.

- Any reason (any)
- No other popular shows on at that time (bored)
- Critics still give the show good reviews (critics)
- Other people still watch the show (peers)
- The original screenwriters stay (writers)
- The original directors stay (director)
- The original cast stays (cast)

Data setup

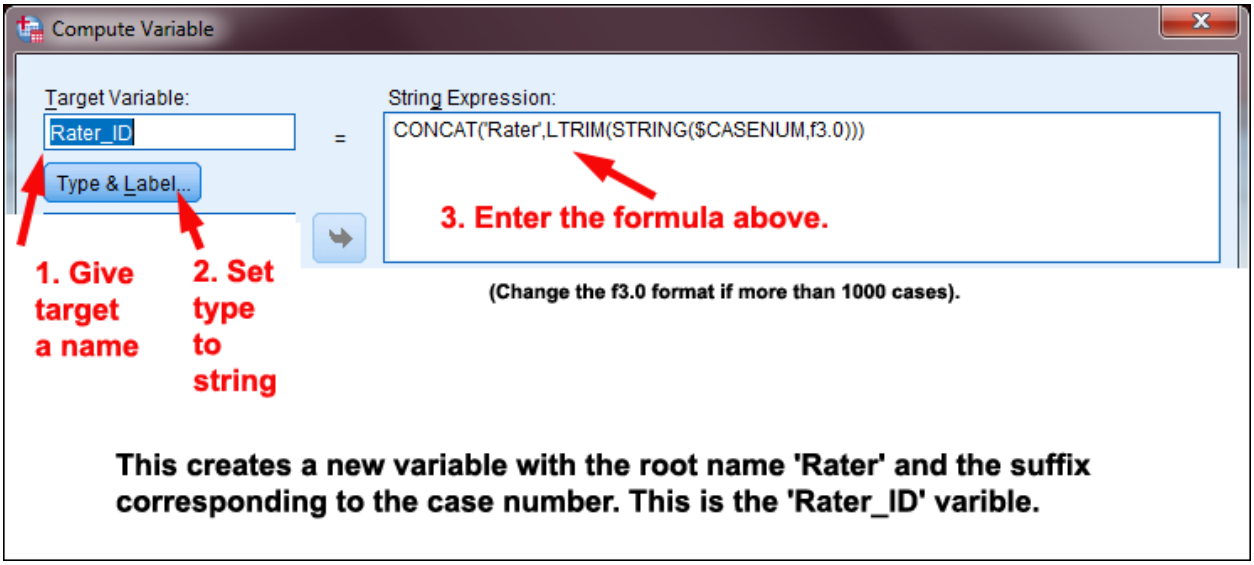
When using intraclass correlation for inter-rater reliability, the researcher constructs a table in which rows are the objects that are rated and columns are the raters or judges. That is, column 1 is the target id (1, 2, ..., n) and subsequent columns are the raters (A, B, C, ...), as shown further below for example data. It may be necessary to transpose the data (Data > Transpose in SPSS menus) to make the raters be the columns, as is explained next for the example data on 906 respondents rating TV shows on 7 items.

Below is the original, untransposed data matrix for the first 10 cases.

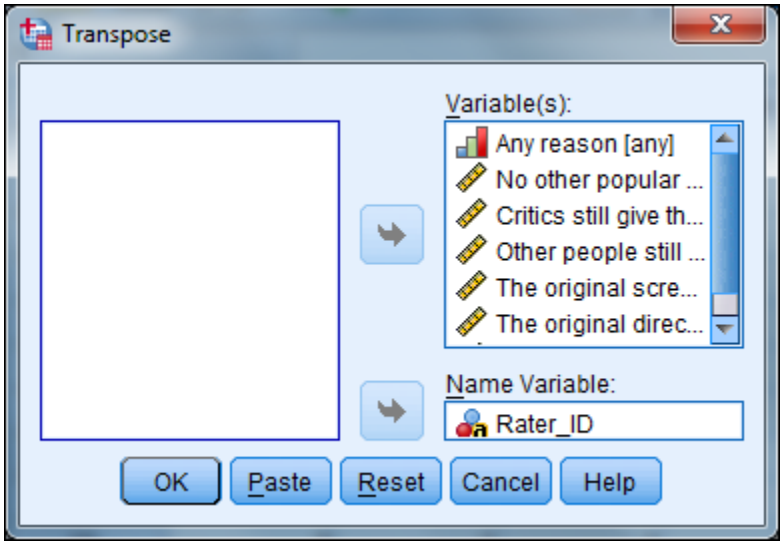


	any	bored	critics	peers	writers	director	cast
1	0	0	0	0	1	1	1
2	1	1	0	1	1	1	1
3	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1
7	0	0	1	1	1	1	1
8	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1
10	0	1	1	1	1	1	1

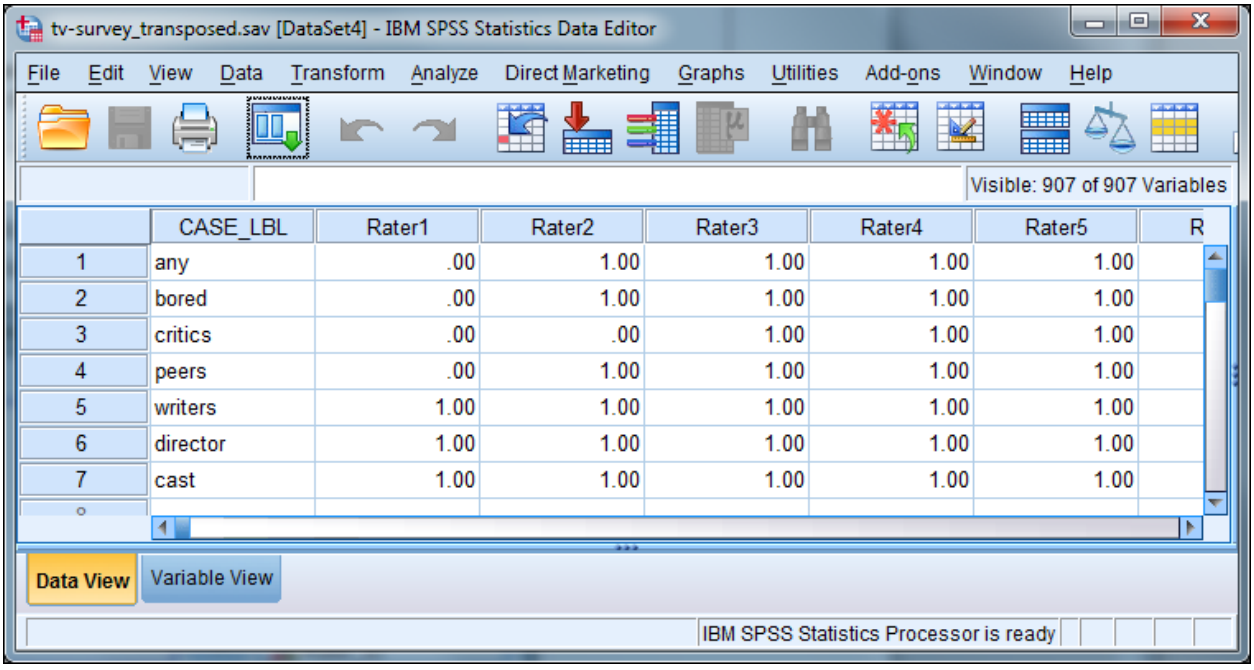
Prior to transposing the data, it is helpful to create a Rater_ID variable, which has the values Rater1, Rater2, Rater3, etc. How to do so in SPSS is shown in the figure below, using Transform > Compute Variable.



The data are transposed in SPSS with the Data > Transpose menu selection, filling out the "Transpose" dialog as illustrated below. Note the new Rater_ID variable is used as the "Name" variable.



After transposition, the data file looks as shown in the figure below. The rows are target of the ratings, here the television survey items discussed above. The cell entries are the raters' ratings of the target on some interval variable or interval-like variable, such as some Likert scale, or, in this example, a binary 0/1 scale. The purpose of ICC is to assess the inter-rater (column) effect in relation to the grouping (row) effect, using two-way ANOVA.



	CASE_LBL	Rater1	Rater2	Rater3	Rater4	Rater5	R
1	any	.00	1.00	1.00	1.00	1.00	
2	bored	.00	1.00	1.00	1.00	1.00	
3	critics	.00	.00	1.00	1.00	1.00	
4	peers	.00	1.00	1.00	1.00	1.00	
5	writers	1.00	1.00	1.00	1.00	1.00	
6	director	1.00	1.00	1.00	1.00	1.00	
7	cast	1.00	1.00	1.00	1.00	1.00	

ICC models and types

ICC varies depending the model and the type.

1. Model: Whether the judges are all judges of interest or are conceived as a random sample of possible judges, and whether all targets are rated or only a random sample. Three types of models are shown below for SPSS output.
2. Type: Whether absolute agreement is required or merely correlational consistency

In SPSS, as illustrated above, there are three models and two types, giving rise to six forms of ICC, as described in the classic article by Shrout and Fleiss (1979) and discussed by McGraw & Wong (1996), whose article is the prima4ry citation in SPSS documentation regarding ICC. Models and types are discussed in more detail below.

MODELS

For all three model types below, it is assumed that the objects being rated represent a random selection of all cases. In SPSS parlance, objects are the "people" factor because it is often people who are the objects of ratings (ex., athletes being rated at the Olympics). Also in SPSS parlance, the columns represent the different "measures" being applied, such as the assessments of different raters. As SPSS terminology can be confusing (ex., raters are indeed people, but so are athletes, but only the athletes-as-objects represent the "people" factor in SPSS), here we refer to the columns as the raters, who are the judges, and we refer to the rows as the objects.

One-way random effects model. This model type is rare because it assumes each object (case, based on rows) is rated by a different randomly selected rater (or different set of randomly selected raters). This model applies even when the researcher cannot associate a particular subject with a particular rater because information is lacking about which judge assigned which score to a subject. SPSS table notation states "People effects are random," meaning object effects are random, as they are in all models. Rater effects are absorbed into error or residual variance. Called "consistency ICC", ICC for the one-way random effects model represents the percent of variance attributable to differences in the

objects. Put another way, ICC is interpreted as the proportion of variance associated with differences among the scores of the subjects/objects/cases.

Two-way random effects model. This is the most common ICC model Judges are conceived as being a random selection from among all possible judges. Raters rate all objects assumed to reflect a random selection from a pool of all possible objects. All objects are rated by all judges and it is known how each judge rated each subject. The ICC is interpreted as the proportion of object plus rater variance that is associated with differences among the scores of the objects. SPSS table notation states, "Both people effects and measures effects are random," meaning that object and rater effects respectively are assumed random. The ICC is interpreted as being generalizable to all possible judges.

Two-way mixed model. This model gives identical computational results for ICC as the two-way random effects model, but is interpreted differently. As for all models, objects are assumed to reflect a random sample. The raters, however, are assumed not to be a random sample but rather to represent a fixed group of raters. SPSS table notation states, "People effects are random and measures effects are fixed," meaning that objects are assumed random but raters are assumed fixed. Because of this the resulting ICC is interpreted as not being generalizable to any other set of raters.

TYPES

Absolute agreement: Measures if raters assign the same absolute score. Absolute agreement is often used when systematic variability due to raters is relevant.

Consistency: Measures if raters' scores are highly correlated even if they are not identical in absolute terms That is, raters are consistent as long as their relative ratings are similar. Consistency agreement is used when systematic variability due to raters is irrelevant.

Single versus average ICC

Each model may have two versions of the intraclass correlation coefficient:

Single measure reliability: Individual ratings constitute the unit of analysis. Single measure reliability gives the reliability for a single judge's rating. In the example

discussed in the SPSS section below, single measure ICC is .143 (for the one-way random effects model), indicating a relatively low level of inter-rater consistency on the seven attribute items taken one at a time. Single measure reliability is the one of usual interest, used to assess if the ratings of one judge are apt to be the same as for another judge. When single measure ICC is low, the researcher must assume different people rate attributes differently.

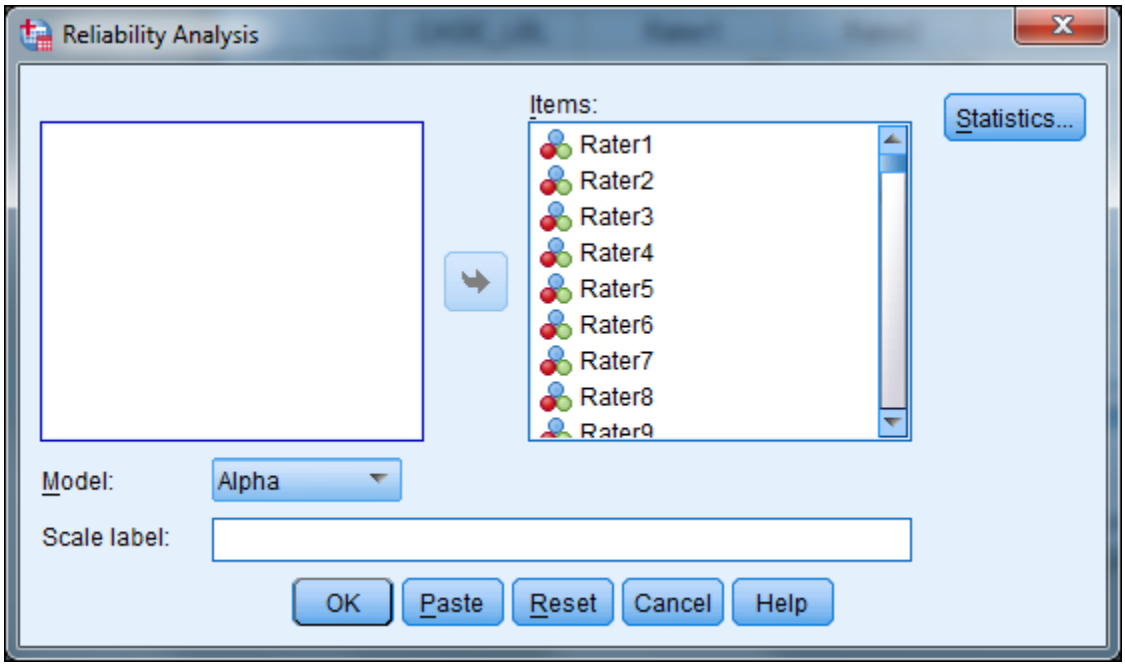
Average measure reliability: The mean of all ratings is the unit of analysis. That is, average measure reliability gives the reliability of the mean of the ratings of all raters. Use this if the research design involves averaging multiple ratings for each item across all raters, perhaps because the researcher judges that using an individual rating would involve too much uncertainty. In the SPSS example below, average measure ICC is .993, indicating a high level of inter-rater consistency on the average of all 7 ratings. That average measure ICC is high means that when attributes are averaged across all raters, the mean ratings are very stable. It does not mean that raters agree in their ratings of individual items (that is single measure ICC), only that they agree in their mean ratings across all items.

Average measure reliability is close to Cronbach's alpha. Average measure reliability for either two-way random effects or two-way mixed models will be the same as Cronbach's alpha. In this example, for the one-way random model, the ICC and Cronbach's alpha differ, but not greatly.

Average measure reliability requires a reasonable number of judges to form a stable average. The number of judges required is estimated beforehand as nj = ICC*(1 - rl)/rl(1 - ICC*), where nj is the number of judges needed, rl is the lower bound from the (1-a)*100% confidence interval around the ICC, discovered in a pilot study; and ICC* is the minimum level of ICC acceptable to the researcher (ex., .80).

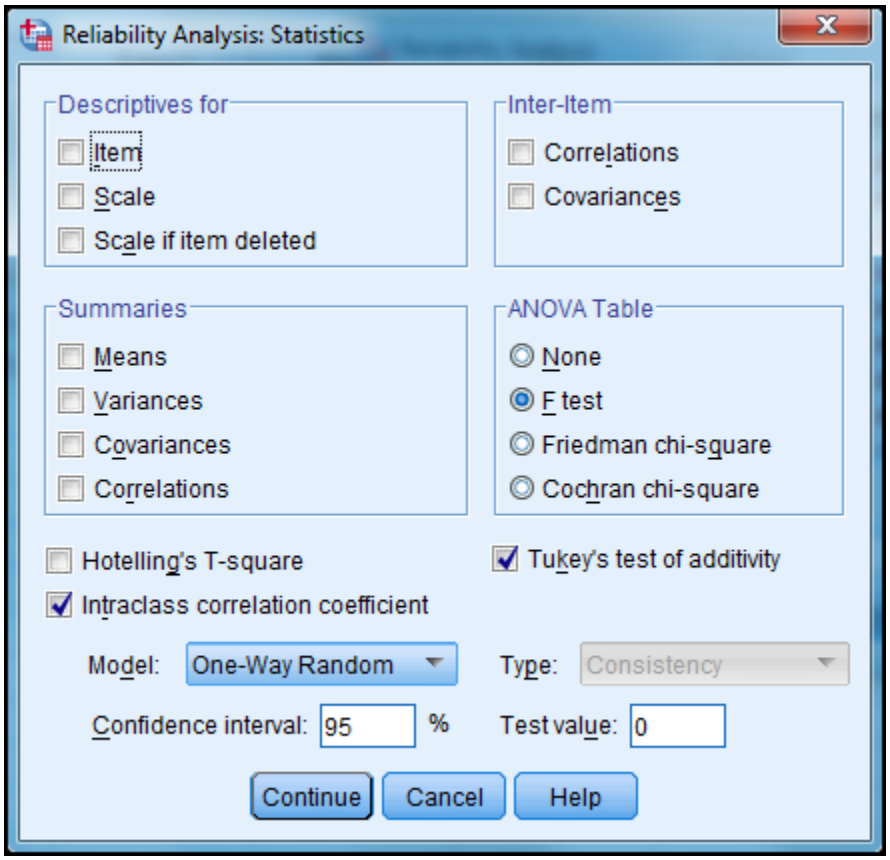
ICC in SPSS

In SPSS, select Analyze > Scale > Reliability Analysis and select the variables, as illustrated below. Since the data are transposed, the raters are the “items” or “variables” which form the columns in the data matrix. The rows are the objects being rated. Though here the objects are television shows, often the objects are people (ex., athletes being rated 0 – 10 in the Olympics) and hence in SPSS literature, somewhat confusingly the objects may be referred to as “persons”.



Then click the Statistics button, shown below. In the "Descriptives" group, select "Item"; optionally in the "ANOVA Table" group, select "F-test"; and near the bottom select "Intraclass correlation coefficient". Warning: if items in the "Summaries" or "Inter-Item" groups are selected and if some raters have no variance (they rated all objects the same), those raters will be dropped from the analysis even for the ICC calculation, which otherwise can tolerate absence of variance.

Select a model from the "Model" drop-down list, (here it is "One-Way Random" but all three selections are shown in output below). Select a type from the "Type" drop-down list (also discussed below). For this example, "Consistency" is chosen as the type. Click Continue, OK.



SPSS ICC output

The intraclass correlation (ICC) is an ANOVA-based measure. To illustrate, below is the ANOVA table for the one-way random effects model for this example. Single-item ICC, which is the usual ICC variant for measuring inter-rater reliability is shown equal to the "Between People" variance divided by "Total" variance, for the "One-Way Random Effects" model. In SPSS parlance, "People" refers not to the raters but to the objects (cases, rows) being rated, who are often people, as in athletic events (The same calculation is not that applicable to other models).

ANOVA with Tukey's Test for Nonadditivity

			Sum of Squares	df	Mean Square	F	Sig
Between People			181.487	6	30.248		
Within People	Between Items		780.866	905	.863	9.762	.000
	Residual	Nonadditivity	64.703[a]	1	64.703	845.945	.000
		Balance	415.239	5429	.076		
		Total	479.942	5430	.088		
	Total		1260.808	6335	.199		
Total			1442.295	6341	.227		

Grand Mean = .6503

a. Tukey's estimate of power to which observations must be raised to achieve additivity = 2.107.

181.487/1260.808 = .143

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.143	.064	.448	151.982	6	6335	.000
Average Measures	.993	.984	.999	151.982	6	6335	.000

One-way random effects model where people effects are random.

Output for the "Two-Way Random Effects" and "Two-Way Mixed Effects" models is shown below. As can be seen, for all three models the "Average Measures" ICC is similar and high, indicating that mean ratings are the same across raters. The "Single Measures" intraclass correlation is always lower, as ratings of single objects will always vary more than will raters' mean scores. This is true under any of the three model assumptions about effects.

Cronbach's alpha, also shown below, is identical regardless of model. However, note that with the untransposed data and the reasons-to-watch-television items as variables, Cronbach's alpha would have its normal function as a reliability criterion for assessing if items cohered enough (ex., > .70) to constitute a scale. Here, with transposed data and raters as variables, Cronbach's alpha is a measure of the reliability of the raters collectively. In the "Reliability Statistics" box, "Items" refers to columns, which are the 906 raters.

Reliability Statistics

Cronbach's Alpha	N of Items
.997	906

The two-way models are type = consistency. See table note b.

ONE-WAY RANDOM EFFECTS MODEL

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.143	.064	.448	151.982	6	6335	.000
Average Measures	.993	.984	.999	151.982	6	6335	.000

One-way random effects model where people effects are random.

TWO-WAY RANDOM EFFECTS MODEL

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.274[a]	.135	.647	342.219	6	5430	.000
Average Measures	.997	.993	.999	342.219	6	5430	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

TWO-WAY MIXED EFFECTS MODEL

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.274[a]	.135	.647	342.219	6	5430	.000
Average Measures	.997[c]	.993	.999	342.219	6	5430	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

All models are significant by the F test, meaning that the ICC differs from 0 under any model. The upper and lower confidence bounds for the single measure ICC are quite broad, indicating the range in which 95% of ICC estimates would fall if additional random samples of raters were selected and employed.

Note that ICC estimates are always the same for the "Two-Way Random" and "Two-Way Mixed" models. It is just that ICC is interpreted differently according to the different model assumptions:

- "Two-Way Random": Raters are assumed to be randomly reflective of all raters and ICC can be generalized to other sets of raters.
- "Two-Way Mixed": Raters are assumed to be a fixed set and the ICC cannot be generalized to another set of raters.

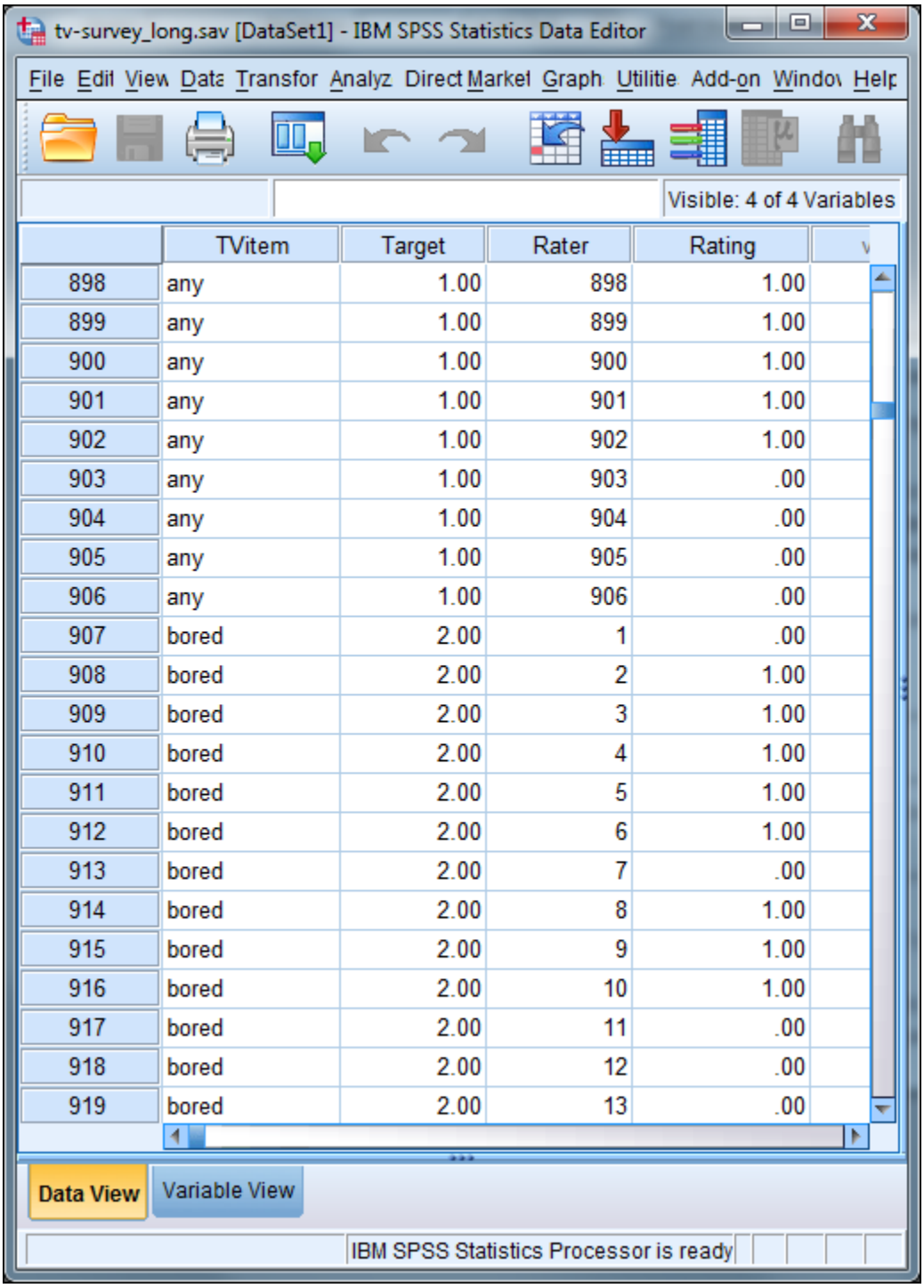
ICC in SAS

SAS does not support a procedure for calculating the intraclass correlation coefficient for the case of multiple raters rating multiple objects. However, a web search for "icc sas macro" will reveal a number of third-party SAS macros for this purpose. One is the "AgreeStat" macro from Advanced Analytics, at the url http://agreestat.com/agreestat_sas.html.

ICC in Stata

Data setup

The Stata example uses the same data as described above for ICC using SPSS. Stata requires that data be in long form for ICC. Long form is illustrated below. Each rated object (here, TVitem) has as many rows as there are raters (906 in this example). As Stata wants the target object being rated to be numeric, a new variable named "Target" was created, such that "any" = 1, "bored" = 2, etc., for each of the seven TVitem variables. A "Rater" variable was also created from the case number in the original dataset, but that variable is not used by the ICC command. Creation of long form data was accomplished using the SPSS Data > Restructure operation but could also have been accomplished with Stata's "reshape" command.



	TVitem	Target	Rater	Rating
898	any	1.00	898	1.00
899	any	1.00	899	1.00
900	any	1.00	900	1.00
901	any	1.00	901	1.00
902	any	1.00	902	1.00
903	any	1.00	903	.00
904	any	1.00	904	.00
905	any	1.00	905	.00
906	any	1.00	906	.00
907	bored	2.00	1	.00
908	bored	2.00	2	1.00
909	bored	2.00	3	1.00
910	bored	2.00	4	1.00
911	bored	2.00	5	1.00
912	bored	2.00	6	1.00
913	bored	2.00	7	.00
914	bored	2.00	8	1.00
915	bored	2.00	9	1.00
916	bored	2.00	10	1.00
917	bored	2.00	11	.00
918	bored	2.00	12	.00
919	bored	2.00	13	.00

Types and Models in Stata

Types (consistent, absolute) and models (one-way random, two-way random, two-way mixed) are available in Stata to generate the same six flavors of the ICC

coefficient as described by Fornell & Larcker (1981) and as illustrated above for SPSS previously. Models and types of ICC are discussed above.

Stata syntax

The "use" command for the one-way random effects model discussed above in the SPSS section invokes the long form version of the TV survey data:

```
. use "C:\Data\tv-survey_long.dta", clear
```

The "icc" command for one-way random effects models takes the form "icc dependentvariable targetvariable":

```
. icc Rating Target
```

The "icc" command for two-way random effects models takes the form "icc dependentvariable targetvariable judgevariable":

```
. icc Rating Target Rater
```

The "icc" command for two-way mixed effects models takes the form "icc dependentvariable targetvariable judgevariable, mixed":

```
. icc Rating Target Rater, mixed
```

By default, the two-way models are absolute in type. To correspond to SPSS output above, type was set to consistent. This is done simply by adding "consistent" (or "absolute") as an option after the comma:

```
. icc Rating Target Rater, mixed consistent
```

Stata output for one-way random effects ICC

Stata output returns the same ICC coefficients and F test significance and has the same interpretation as for the one-way random effects ICC model discussed above.

```
. use "C:\Data\tv-survey_long.dta", clear

. icc Rating Target

Intraclass correlations
One-way random-effects model
Absolute agreement

Random effects: Target              Number of targets =          7
                                    Number of raters  =        906

------------------------------------------------------------------------------
                Rating |        ICC       [95% Conf. Interval]
-----------------------+------------------------------------------------------
            Individual |   .1428422        .0641049    .4482788
               Average |   .9934203        .9841414    .9986434
------------------------------------------------------------------------------
F test that
  ICC=0.00: F(6.0, 6335.0) = 151.98           Prob > F = 0.000

Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

Stata output for two-way random and two-way mixed effects ICC

Parallel output for the two-way random effects model and the two-way mixed model can be obtained using the commands discussed above in the Stata ICC syntax section. The format is identical to that above for the one-way random effects model and ICC coefficients, confidence intervals and the F test is identical to SPSS output shown earlier above.

Stata saved results and parameters

Stata saves many output and parameter variables in memory. These can be displayed with the command "`return list, all`", as illustrated below. The command "`return list`" will return the same list minus hidden variables. The figure below shows the listing after a prior icc command asking for a two-way mixed effects model of absolute type.

```
. return list, all

scalars:
            r(N_target) =  7
             r(N_rater) =  906
               r(icc_i) =  .1432927797242789
             r(icc_i_F) =  342.2192252295826
           r(icc_i_df1) =  6
           r(icc_i_df2) =  5430
             r(icc_i_p) =  0
            r(icc_i_lb) =  .0645497035575729
            r(icc_i_ub) =  .4485647456397424
             r(icc_avg) =  .993444234424065
           r(icc_avg_F) =  342.2192252295826
         r(icc_avg_df1) =  6
         r(icc_avg_df2) =  5430
           r(icc_avg_p) =  0
          r(icc_avg_lb) =  .9842563119578696
          r(icc_avg_ub) =  .998644959403228
           r(testvalue) =  0
               r(level) =  95

macros:
              r(model) : "two-way mixed effects"
              r(depvar) : "Rating"
              r(target) : "Target"
               r(rater) : "Rater"
                r(type) : "absolute"

Hidden; names and contents could change in future versions of Stata

scalars:
                  r(bms) =  30.24776621465367
                  r(jms) =  .8628353291483072
                  r(wms) =  .1990225646440201
                  r(ems) =  .0883871038933056
```

Reliability: Assumptions

Ordinal or interval items of similar difficulty

Internal consistency analysis using Cronbach's alpha assumes the scale items are ordinal or interval and all measure the same dimension equally (ex., and assortment of math problems of equal difficulty). That is, the reliability methods outlined in this volume assume items comprise an ordinal or interval scale. Cronbach's alpha and other reliability measures discussed in this volume are not appropriate for ordered scales (ex., Guttman scales, where the scale score is passing the most difficult item, based on more difficult items predicting the responses to any less difficult items) or for composite scores (the sum of items reflects the desired construct but items do not necessarily correlate highly, as in a scale of philanthropy where items are dollars given to various types of causes). For a discussion of indexes, ordinal scales, ordered scales, and composite scores, see the separate Statistical Associates "Blue Book" volume on "Scales and Measures."

Additivity

Each item should be linearly related to the total score. *Tukey's test of non-additivity*, a choice under the Statistics button of the Reliability dialog in SPSS, discussed above, tests the null hypothesis that there is no multiplicative interaction between the cases and the items. If this test is significant (<= .05) then there is multiplicative interaction. The Tukey significance is found in the "Nonadditivity" row of the "ANOVA with Tukey's Test for Nonadditivity" table in SPSS output.

If Tukey's test shows multiplicative interaction, any model computing scores for cases based on the scale must include the case main effect, the item main effect, and the case-by-item interaction effect. In a footnote to the Tukey test output, SPSS prints an estimate of the power to which items in a set would need to be raised in order to be additive. (Warning: while transforms may eliminate non-additivity, raising item scores to too high a power will generate large values for all subjects, obscuring differences among subjects).

In SPSS, select Analyze, Scale, Reliability Analysis; click Statistics; check Tukey's test of additivity. The output below is General Social Survey data for the four

education items in illustrations above. Since Tukey's test is significant, multiplicative interaction is indicated for these data.

ANOVA with Tukey's Test for Nonadditivity

			Sum of Squares	df	Mean Square	F	Sig
Between People			7785.301	1123	6.933		
Within People	Between Items		125804.641	3	41934.880	20674.086	.000
	Residual	Nonadditivity	3616.005[a]	1	3616.005	3785.022	.000
		Balance	3217.604	3368	.955		
		Total	6833.609	3369	2.028		
	Total		132638.250	3372	39.335		
Total			140423.551	4495	31.240		

Grand Mean = 4.17

a. Tukey's estimate of power to which observations must be raised to achieve additivity = .463.

Independence

Observations for one subject/case should be independent of observations for any other subject/case in any administration of the instrument. However, the fact that test-retest designs involve correlated data between administrations does not pose a statistical problem in assessing reliability and does not in itself violate assumptions of reliability analysis.

Uncorrelated error

Errors should be uncorrelated .

Consistent coding

High values must have the same meaning across items.

Random assignment of subjects

In split-half tests, random assignment of items to forms is assumed. Typically, odd-numbered items become one form and even-numbered items become the second form. Sequential assignment may involve a subject fatigue factor with regard to the second form.

Equivalency of forms

In split-half tests (recall split half analysis is selected in the "Model" pull-down of the initial SPSS Reliability dialog), the two forms should be equivalent. A test of this is to see if the mean response is the same in the two groups. In split half models, *Hotelling's* T^2 is a multivariate test for equality of means between groups. A significant T^2 means that the null hypothesis that means were equal can be rejected by the researcher. Apart from split-half test analyses, T^2 tests if the means of a set of items differ and again, a significant T^2 means one rejects the null assumption of equal means, as in the example below for data on the four-item education scale in previous examples. This test assumes multivariate normality of items. Hotelling's T^2 is a choice under the Statistics button of the Reliability dialog in SPSS.

Hotelling's T-Squared Test

Hotelling's T-Squared	F	df1	df2	Sig
39068.394	12999.605	3	1121	.000

Equal variances

In split-half tests, the Spearman-Brown split-half reliability coefficient assumes the split halves have equal variances. The *chi-square test of parallel models* tests the null hypothesis that the variances are equal. If the chi-square significance is <= .05, then the researcher concludes the models are not parallel and that the variances differ significantly. In SPSS, select "Parallel" under the model button.

Same assumptions as for correlation

Reliability analysis involves correlation and as such it requires that linearity, homoscedasticity, and other assumptions common to general linear models.

Reliability: Frequently Asked Questions

How is reliability related to validity?

A measure may be reliable but not valid, but it cannot be valid without being reliable. That is, reliability is a necessary but not sufficient condition for validity. As just one example, in split-half reliability, the two batteries may correlate highly on the Spearman-Brown coefficient, indicating reliability, but when the scale is considered along with other scales in the model, it may lack divergent validity.

How is Cronbach's alpha related to factor analysis?

Both are methods of establishing internal consistency in a set of items constituting a scale. As a rule of thumb, the set of items may be said to constitute a good scale for confirmatory purposes if alpha >= .8, an adequate scale for confirmatory purposes if alpha >=.7, and a scale for exploratory purposes if alpha >= .6. A set of items may be said to constitute a scale in factor analysis if there is simple factor structure: there are no crossloadings in the .3 to .7 range, so all items load heavily on only one factor.

Usually researchers will select either the Cronbach's alpha test of internal consistency or the factor analysis test, not both. Usually both tests will agree. If they do not agree, this may indicate a marginal scale. If choosing between the two tests, one may consider Cronbach's alpha is sensitive to number of items in the scale whereas the factor method is not. Also, Cronbach's alpha is a type of bivariate method whereas factor analysis is multivariate. That is, the Cronbach's alpha numerator is the number of items times the average of the covariances of all pairs of items. The denominator is the average variance of the items plus the quantity N-1 times the average covariance, where N is the number of items. Factor loadings reflect the linear relationship of the loaded indicator variables with the factor or component, controlling for all other variables in the model. The two methods do not define internal consistency in the same manner and may yield different results.

How is reliability related to attenuation in correlation?

Reliability is a form of correlation. Correlation coefficients can be attenuated (misleadingly low) for a variety of reasons, including truncation of the range of

variables (as by dichotomization of continuous data; reducing a 7-point scale to a 3-point scale). Measurement error also attenuates correlation. Reliability may be thought of as the correlation of a variable with itself. Attenuation-corrected correlation ("disattenuated correlation") is higher than the raw correlation on the assumption that the lower the reliability, the greater the measurement error, and the higher the "true" correlation is in relation to the measured correlation.

The *Spearman correction for attenuation* of a correlation: let r_{xy}* be corrected r for the correlation of x and y; let r_{xy} be the uncorrected correlation; then r_{xy}* is a function of the reliabilities of the two variables, r_{xx} and r_{yy}:

$$r_{xy}^* = r_{xy} / [SQRT\{r_{xx}r_{yy}\}]$$

This formula will result in an estimated true correlation (r_{xy}*) which is higher than the observed correlation (r_{xy}), and all the more so the lower the reliabilities. Corrected r may be greater than 1.0, in which case it is customarily rounded down to 1.0.

Note that use of attenuation-corrected correlation is the subject of controversy (see, for ex., Winne & Belfry, 1982). Moreover, because corrected r will no longer have the same sampling distribution as r, a conservative approach is to take the upper and lower confidence limits of r and compute corrected r for both, giving a range of attenuation-corrected values for r. However, Muchinsky (1996) has noted that attenuation-corrected reliabilities, being not directly comparable with uncorrected correlation, are therefore not appropriate for use with inferential statistics in hypothesis testing and this would include taking confidence limits. Still, Muchinsky and others acknowledge that the difference between a correlation and attenuation-corrected correlation may be useful, at least for exploratory purposes, in assessing whether a low correlation is low because of unreliability of the measures or because the measures are actually uncorrelated.

How should a negative reliability coefficient be interpreted?

It should be interpreted as a data entry error, a data measurement problem, as a problem based on small sample size, or as indicative of multidimensionality. As negative reliability is rare, the researcher should first check to see if there are coding or data entry errors.

One situation in which negative reliability might occur is when the scale items represent more than one dimension of meaning, and these dimensions are negatively correlated, and one split half test is more representative of one dimension while the other split half is more representative of another dimension. As Krus & Helmstadter point out, factor analyzing the entire set of items first would reveal if the set of items is plausibly conceptualized as unidimensional.

A second scenario for negative reliability is discussed by Magnusson (1966: 67), who notes that when true reliability approaches zero and sample size is small, random disturbance in the data may yield a small negative reliability coefficient.

In the case of Cronbach's alpha, Nichols (1999) notes that values less than 0 or greater than 1.0 may occur, especially when the number of cases and/or items is small. Negative alpha indicates negative average covariance among items, and when sample size is small, misleading samples and/or measurement error may generate a negative rather than positive average covariance. The more the items measure different rather than the same dimension, the greater the possibility of negative average covariance among items and hence negative alpha.

What is Cochran's Q test of equality of proportions for dichotomous items?

Cochran's Q is used to test whether a set of dichotomous items split similarly. This is the same as testing whether the items have the same mean. If they test the same, then items within the set might be substituted for one another. In the ANOVA output table for a set of dichotomous items, the "Between Items" row, "Sig" for Cochran's Q column, if Sig (Q) <= .05, then the researcher rejects the null hypothesis that all items display an equal split or have the same mean.

In SPSS, select Analyze, Scale/Reliability; select your items; click Statistics; in the Descriptives area, select Item, Scale, Scale if Deleted; in Summarize, select summary statistics (Means, Variances, Covariances, Correlations); and in the ANOVA table group, select Cochran chi-square. Continue. OK.

Cochran's Q is discussed further in the separate Statistical Associates "Blue Book" volume on measures of association.

What is the derivation of intraclass correlation coefficients?

Derivation of the ICC formula, following Ebel (1951: 409-411): Let A be the true variance in subjects' ratings due to the normal expectation that different subjects will have true different scores on the rating variable. Let B be the error variance in subjects' ratings attributable to inter-rater unreliability. The intent of ICC is to form the ratio, ICC = A/(A + B). That is, intraclass correlation is to be true inter-subject variance as a percent of total variance, where total variance is true variance plus variance attributable to inter-rater error in classification. B is simply the mean-square estimate of within-subjects variance (variance in the ratings for a given subject by a group of raters), computed in ANOVA. The mean-square estimate of between-subjects variance equals k times A (the true component) plus B (the inter-rater error component), since each mean contains a true component and an error component.

Given B = ms_{within}, and given $ms_{between}$ = kA + B, substituting these equalities into the intended equation (ICC = A/[A+B]), the equation for ICC reduces to the formula for the most-used version of intraclass correlation (Haggard, 1958: 60):

$ICC = r_I = (ms_{between} - ms_{within})/(ms_{between} + [k - 1]ms_{within})$

where

$ms_{between}$ is the mean-square estimate of between-subjects variance, reflecting the normal expectation that different subjects will have true different scores on the rating variable

ms_{within} is the mean-square estimate of within-subjects variance, or error attributed to inter-rater unreliability in rating the same person or target (row).

k is the number of raters/ratings per target (person, neighborhood, etc.) = number of columns. If the number of raters differs per target, an average k is used based on the harmonic mean: $k' = (1/[n-1])(SUMk-[SUMk^2]/SUMk)$.

What are Method 1 and Method 2 in the SPSS RELIABILITY module?

These are two methods of computing the reliability coefficient (Cronbach's alpha) for a set of items thought to comprise a scale. Method 1 allows constant terms to

remain in the scale, while Method 2 deletes constant terms. Method 2 also produces a standardized item alpha, as if data had been input in standardized form. Method 2 can be forced when using the syntax window by adding the clause METHOD=COV. Method 2 is the default in SPSS.

Acknowledgments

Appreciation is expressed to peer reviewers of an earlier draft of this manuscript for their generous donation of time and effort, by Professor Immanuel Thomas, Chairman, Department of Psychology, Kerala University, India; and Dr. Jolita Bekhof, Faculty, Department of Pediatrics, Isala Klinieken, Netherlands.

Bibliography

Adcock, R. and D. Collier (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95: 529-546.

APA (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51(2, supplement): 201-238.

APA (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

Armor, D. J. (1974). Theta reliability and factor scaling. Pp. 17-50 in H. Costner, ed., *Sociological methodology*. San Francisco: Jossey-Bass.

Bagozzi, R. P., Y. Yi and L. W. Phillips (1991). Assessing Construct Validity in Organizational Research. *Administrative Science Quarterly* 36(3): 421-458.

Bollen, Kenneth A. & Ward, Sally (1979). Ratio variables in aggregate data analysis: Their uses, problems, and alternatives. *Sociological Methods & Research* 7(4): 431-450.

Bonett, Douglas G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine* 21: 1331-1335.

Campbell, Donald T. & J. C. Stanley (1963a). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage, ed., *Handbook of research on teaching*. Chicago: Rand McNally, 1963: 171-246 . The seminal article on types of validity.

Campbell, Donald T. & Stanley, J. C. (1963b). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment*. Quantitative Applications in the Social Sciences series 07-017. Newbury Park, CA: Sage Publications.

Chin, W. W. (1998). *The partial least squares approach for structural equation modeling.* Pp. 295-336 in G. A. Macoulides, ed. *Modern methods for business research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (ed.). *Educational measurement. Second ed.*. Washington, D. C.: American Council on Education.

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

Cook, Thomas D. & Campbell, Donald T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin. Chapter 2 is a classic statement of types of validity.

Daskalakis, Stylianos & Mantas, John (2008). Evaluating the impact of a service-oriented framework for healthcare interoperability. Pp. 285-290 in Anderson, Stig Kjaer; Klein, Gunnar O.; Schulz, Stefan; Aarts, Jos; & Mazzoleni, M. Cristina, eds. *eHealth beyond the horizon - get IT there: Proceedings of MIE2008* (Studies in Health Technology and Informatics). Amsterdam, Netherlands: IOS Press, 2008.

Ebel, Robert L. (1951). Estimation of the reliability of ratings. *Psychometrika* 16: 407-424.

Fleiss, J. L., Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33: 613-619.

Fornell, C., & Larcker, F. D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1): 39-50.

Gibbons, J. D. (1985). *Nonparametric statistical inference.* New York: Marcel Dekker.

Graham, James M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them: *Educational and Psychological Measurement* 66; 930-944.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6(5): 427-439.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance*. NY: Dryden.

Hair, J.; Black, W.; Babin, B.;,& Anderson, R. (2010). *Multivariate data analysis, Seventh ed.* Upper Saddle River, NJ: Prentice-Hall.

Krus, D.J., & Helmstadter, G.C. (1993) The problem of negative reliabilities. *Educational and Psychological Measurement*, 53, 643-650.

Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33:159-174. This article sets cut-offs for Cohen's Kappa.

Litwin, Mark S.(2002). *How to assess and interpret survey psychometrics*. The Survey Kit series, Vol. 8. Thousand Oaks, CA: Sage Publications). Covers test-retest, alternate-form, internal consistency, interobserver, and intraobserver reliability.

Magnusson, D. (1966). *Test theory.* Reading, MA: Addison-Wesley.

McGraw, K. O. and S.P. Wong (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1(1): 30-46. See also "Correction", *Psychological Methods* 1(4): 390.

McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? *Journal of Gen Psychology* 119(1):59-72. This article reports that reliability estimates under test-retest designs are not inflated due to memory effects.

McNemar, Q. (1969). *Psychological Statistics. Fourth edition*. New York: Wiley Covers F tests for intraclass correlation (p. 322).

Messick, S. (1989). Validity. Pp. 13-103 in R. Linn, ed., *Educational measurement. Third ed.* New York: American Council on Education and Macmillan Publishing Company.

Muchinsky P.M. (1996) The correction for attenuation. *Educational & Psychological Measurement* 56(1), 63-75.

Myers, L. S.; Gamst, G.; & Guarineo, A. J. (2013). *Applied Multivariate Research, Second Ed*. Thousand Oaks, CA: Sage Publications.

Nichols, David P. (1999). My coefficient a is negative! *SPSS Keywords*, Number 68. Retrieved 1/28/2009 from http://www.ats.ucla.edu/stat/Spss/library/negalpha.htm.

Nunnelly, J. C. (1970) *Psychometric Theory*. *Second ed*., 1978. New York: McGraw Hill. Cited as a reference in support of the .70 cut-off for Cronbach's alpha. Classic on reliability in psychological and educational testing.

Podsakoff, MacKenzie, Lee, & Podsakoff. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5): 879-903.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Raykov, Tenko (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173-184.

Raykov, Tenko (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items *Applied Psychological Measurement*, 22(4), 375-385.

Ringle, Christian (2006). Segmentation for path models and unobserved heterogeneity: The finite mixture partial least squares approach. Hamburg, Germany: *Research Papers on Marketing and Retailing*, No. 35, University of Hamburg.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston:Houghton Mifflin.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond, ed., Review of research in education, Vol. 19, pp. 405-450. Washington, DC: American Educational Research Association.

Shrout, P.E., and J. L. Fleiss (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* (86): 420-428. Classic article on intraclass correlation.

SPSS (1988). *SPSS-X User's Guide, Third ed.*. Chicago, IL: SPSS Inc.

Straub, D.W. (1989). "Validating Instruments in MIS Research," *MIS Quarterly*, 13(2): 147-166.

Tenvergert, E., Gillespie, M., & Kingma, J. (1993). Testing the assumptions and interpreting the results of the Rasch model using log-linear procedures in SPSS. *Behavior Research Methods, Instruments & Computers* 25(3), 350-359.

Walter, S. D.; Eliasziw, M.; and Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine* 17: 101-110.

Winne, Philip H. & Belfry, M. Joan (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, 19(2), 125-134.

Wright B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement* 14, 97–116.

Zumbo, B. D.; Gadermann, A. M.; & Zeisser, C.. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, 6, 21-29.

Wright, B. D. (1996) Comparing Rasch measurement with factor analysis *Structural Equation Modeling* 3, 3-24.

Statistical Associates Publishing
Blue Book Series

Assumptions, Testing of
Canonical Correlation
Case Studies
Cluster Analysis
Correlation
Correlation, Partial
Correspondence Analysis
Cox Regression
Creating Simulated Datasets
Crosstabulation
Curve Estimation and Nonlinear Regression
Delphi Method in Quantitative Research
Discriminant Function Analysis
Ethnographic Research
Factor Analysis
Game Theory
Generalized Linear Models/Generalized Estimating Equations
GLM (Multivariate), MANOVA, and MANCOVA
GLM (Univariate), ANOVA, and ANCOVA
Grounded Theory
Hierarchical Linear Modeling/Multilevel Analysis/Linear Mixed Models
Life Tables and Kaplan-Meier Survival Analysis
Logistic Regression
Log-linear Models,
Longitudinal Analysis
Missing Values Analysis
Multidimensional Scaling
Multiple Regression
Narrative Analysis
Network Analysis
Neural Network Models
Ordinal Regression
Parametric Survival Analysis
Partial Least Squares Regression
Participant Observation
Path Analysis
Power Analysis
Probability
Probit and Logit Response Models

Research Design
Scales and Measures
Significance Testing
Structural Equation Modeling
Survey Research and Sampling
Two-Stage Least Squares Regression
Validity and Reliability
Weighted Least Squares Regression

Statistical Associates Publishing
http://www.statisticalassociates.com
sa.publishers@gmail.com