Internal Consistency Reliability

It is very common in psychological research to collect multiple measures of the same construct. For example, in a questionnaire designed to measure optimism, there are typically many items that collectively measure the construct of optimism. To have confidence in a measure such as this, we need to test its **reliability**: the degree to which it is error-free. The type of reliability we'll be examining here is called **internal consistency reliability**: the degree to which multiple measures of the same thing agree with one another.

Benevolent Sexism Scale

Peter Glick and Susan Fiske (1996) developed an interesting measure called the Benevolent Sexism Scale (BSS). Its 11 items are given below:

1.	No matter how accomplished he is, a man is not truly complete as a person unless he has the love of a woman.
2.	In a disaster, women ought not necessarily to be rescued before men.
3.	People are often truly happy in life without being romantically involved with a member of the other sex.
4.	Many women have a quality of purity that few men possess.
5.	Women should be cherished and protected by men.
6.	Every man ought to have a woman whom he adores.
7.	Men are complete without women.
8.	A good woman should be set on a pedestal by her man.
9.	Women, compared to men, tend to have a superior moral sensibility.
10.	Men should be willing to sacrifice their own well being in order to provide financially for the women in their lives.
11.	Women, as compared to men, tend to have a more refined sense of culture and good taste.

Reverse Scoring

Most of the items above are phrased so that strong agreement indicates a belief that men should protect women, that men need women, and that women have positive qualities that men lack. However, three of the items are phrased in the reverse: #2, #3, and #7. In order to make those items comparable to the other items, we will need to reverse score them.

In this questionnaire, participants responded to the items using a 7-point Likert scale ranging from 1 ("Strongly Disagree") to 7 ("Strongly Agree"). When we reverse-score an item, we want 1's to turn into 7's, 7's to turn into 1's, and all the scores in between to become their appropriate opposite (6's into 2's, 5's into 3's, etc.). Fortunately, there is a simple mathematical rule for reverse-scoring:

$$\text{reverse score}(x) = \max(x) + 1 - x$$

Where max(x) is the maximum possible value for x. In our case, max(x) is 7 because the Likert scale only went up to 7. To reverse score, we take 7 + 1 = 8, and subtract our scores from that. 8 - 7 = 1, 8 - 1 = 7. Voila.

See the Statistics Assignment on reliability (online) for instructions on reverse-scoring variables in SPSS.

Interpreting Reliability Output from SPSS

The following is the output from a reliability analysis on the Benevolent Sexism Scale, with items 2, 3, and 7 reverse-scored.

```
****** Method 1 (space saver) will be used for this analysis ******

  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)


Item-total Statistics

               Scale          Scale      Corrected
               Mean         Variance       Item-           Alpha
              if Item       if Item       Total          if Item
              Deleted       Deleted    Correlation       Deleted

BSS01         42.7568       67.1729        .5980           .6901
BSS04         43.6892       72.9843        .4337           .7160
BSS05         41.8108       80.5665        .2547           .7374
BSS06         42.2973       69.9926        .5137           .7039
BSS08         43.2838       73.7403        .3848           .7227
BSS09         43.9324       74.0365        .4109           .7192
BSS10         43.0541       77.8053        .2570           .7394
BSS11         43.3919       74.6799        .3623           .7257
BSS02R        42.7703       77.1657        .2810           .7363
BSS03R        43.2297       73.4670        .3706           .7250
BSS07R        42.4865       74.8286        .3703           .7246

Reliability Coefficients

N of Cases =     74.0                    N of Items = 11

Alpha =    .7412
```

The first two columns generally aren't all that useful. The third column (“Corrected Item-Total Correlation”) is the correlation between a particular item and the sum of the rest of the items. This tells you how well a particular item "goes with" the rest of the items. In the output above, the best item appears to be BSS01, with an item-total correlation of *r* = .5980. The item with the lowest item-total correlation is BSS05 (*r* = .2547). If the item-total correlation is close to zero, then you should consider removing the item from your scale because it is not measuring the same thing as the rest of the items. If the item-total correlation is negative (e.g., *r* = -.24), then you should either reverse-score that item or remove it. An even better guide for improving the reliability of your scale than item-total correlation is to use the “alpha if item deleted” column, discussed below.

Look at the bottom of the output and you will see "Alpha = .7412." This is a measure of **Cronbach's alpha**, which is the most common statistic used to describe the internal consistency reliability of a set of items. If you are using a questionnaire in your research, your results should include a report of the Cronbach's alpha for your questionnaire.

Alpha if Item Deleted

Now look in the last column: "Alpha if item deleted." This is a very important column. It tells you what the Cronbach's alpha would be if you got rid of a particular item. For example, at the very top of this column, the number is .6901. That means that the Cronbach's alpha of this scale would drop from .7412 to .6901 if you got rid of that item. Because a higher alpha indicates more reliability, it would be a bad idea to get rid of the first item. In fact, if you look down the "Alpha if item deleted" column, you will see that none of the

values is greater than the current alpha of the whole scale: .7412. This means that you don't need to drop any items.

Improving Reliability

If you are using an accepted scale obtained from a published source, you do not need to worry about improving reliability. In general, you should use the whole scale, even if it has problems, because if you start changing the scale you will be unable to compare your results to the results of others who have used the scale. You only want to improve the reliability of a scale if it is a scale you are developing.

If one of the "Alpha if item deleted" values is greater than the overall alpha, you can remove the offending item and then re-run the analysis. Repeating this process until there are no values in the "Alpha if item deleted" column greater than the alpha for the overall scale will improve the reliability of your scale. Why not just take out all the “bad” items at the same time? Strangely enough, removing one item will change the fit of the remaining items, sometimes making an originally bad item into a better item.

Inter-rater Reliability

Percent agreement

If you have two judges who are using a nominal-scale dependent variable (e.g., helpful, neutral, hurtful), an estimate of inter-rater reliability can be obtained by counting the number of times that two judges agree (record the same behavior) and dividing it by the number of possible agreements. A problem with percent agreement is that it can be artificially inflated (seem better than it really is) if there are a very large number of possible agreements or only a few levels of the nominal-scale variable. Cohen’s Kappa was developed to address this problem.

Cohen’s Kappa (κ)

Cohen’s Kappa is a measure of inter-rater reliability for two judges who are using a nominal-scale dependent variable. For example, assume that two judges are watching children on the playground and record each interaction as either “helpful,” “neutral,” or “hurtful.” We begin by displaying their codings using the following 3 x 3 table:

		Judge 2			
		Helpful	Neutral	Hurtful	**Row Totals**
Judge 1	Helpful	21	7	3	**31**
	Neutral	4	58	6	**68**
	Hurtful	1	3	12	**16**
	Column Totals	**26**	**68**	**21**	**Grand Total: 115**

A brief look at this table shows us that the largest numbers are on the main diagonal (Helpful-Helpful, Neutral-Neutral, Hurtful-Hurtful), indicating that the judges generally agreed with one another. Percent agreement is computed by summing the agreements (21 + 58 + 12 = 91) and dividing by the total possible (115), for a percent agreement of 91 / 115 = 79%. Cohen’s Kappa begins by computing the *expected* number of agreements, given the row and column totals. To do this, begin with just the row and column totals. To compute the expected number of agreements, take $\frac{RowTotal}{GrandTotal} \times ColumnTotal$. For expected agreement on Helpful-Helpful, this would be $\frac{31}{115} \times 26 = 7.01$. Repeating this for the other two possible agreement combinations gives us:

		Judge 2			
		Helpful	Neutral	Hurtful	**Row Totals**
Judge 1	Helpful	7.01			**31**
	Neutral		40.21		**68**
	Hurtful			2.92	**16**
	Column Totals	**26**	**68**	**21**	**Grand Total: 115**

The sum of the expected agreements is 7.01 + 40.21 + 2.92 = 50.14. The formula for Cohen's Kappa is:

$$\kappa = \frac{\#\,Agreements - \#\,ExpectedAgreements}{GrandTotal - \#\,ExpectedAgreements}$$

For our example, it would be:

$$\frac{91 - 50.14}{115 - 50.14} = \frac{40.86}{64.86} = .63$$

If Cohen's Kappa is less than .7, the judges are not considered to be consistent in their coding of behavior. Ideally, judges should practice coding the same behaviors before the study begins until their reliability is above .7.

Cronbach's alpha

Cronbach's alpha is most commonly used to measure internal consistency reliability, such as the reliability of a 10-item questionnaire. However, it can also be used to measure inter-rater reliability if the judges used an interval- or ratio-scale dependent variable. It also has the advantage of accommodating more than two judges. To use Cronbach's alpha to compute inter-rater reliability, structure your SPSS data file so that the judges are in columns and the stimuli (e.g., ice skaters) being judged are in rows:

	Judge1	Judge2	Judge3
Ice Skater 1	8	8	9
Ice Skater 2	6	7	6
Ice Skater 3	9	9	9

In your analysis, treat the 3 Judge variables as if they were items in a questionnaire. Just as Cronbach's alpha tells you the degree to which multiple measures tend to "go together," it tells you the degree to which multiple observers agree. As with Cohen's Kappa, Cronbach's alpha should be above 0.7 to have acceptable inter-rater reliability.