Handout 4: Establishing the Reliability of a Survey Instrument

STAT 335 – Fall 2016

In this handout, we will discuss different types of and methods for establishing *reliability*. Recall that this concept was defined in the previous handout as follows.

Definition
Reliability is the extent to which repeatedly measuring the same thing produces the same result.

In order for survey results to be useful, the survey must demonstrate **reliability**. The best practices for questionnaire design discussed in the previous handout help to maximize the instrument's reliability.

THEORY OF RELIABILITY

Reliability can be thought of as follows: $\frac{\text{true score variance}}{\text{observed score variance}}$.

In some sense, this is the proportion of "truth" in a measure. For example, if the reliability is estimated to be .5, then about half of the variance of the observed score is attributable to truth; the other half is attributable to error. What do you suppose is the desired value for this quantity?

Note that the denominator of the equation given above can be easily computed. The numerator, however, is unknown. Therefore, we can never really compute reliability; we can, however, *estimate* it. In the remainder of this handout, we will introduce various types of reliability relevant to survey studies and discuss how reliability is estimated in each case.

TYPES AND MEASURES OF RELIABILITY RELEVANT TO SURVEY STUDIES

When designing survey questionnaires, researchers may consider one or more of the following classes of reliability.

Types of Reliability
Test-Retest Reliability – this is used to establish the consistency of a measure from one time to another.
Parallel Forms Reliability – this is used to assess whether two forms of a questionnaire are equivalent.
Internal Consistency Reliability - this is used to assess the consistency of results across items within a single survey instrument.

Each of these is discussed in more detail below.

Test-Retest Reliability

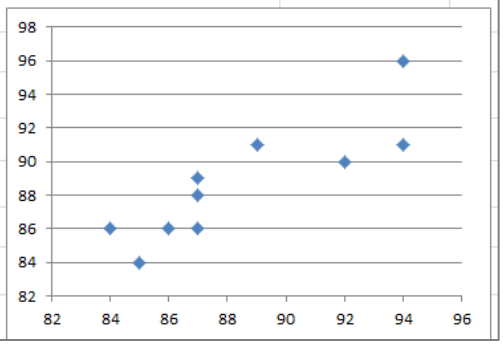
We estimate test-retest reliability when we administer the same questionnaire (or test) to the same set of subjects on two different occasions. Note that this approach assumes there is no substantial change in what is being measured between the two occasions. To maximize the chance that what is being measured is not changing, one shouldn't let too much time pass between the test and the retest.

There are several different measures available for estimating test-retest reliability. In particular, we will discuss the following in this handout:

- Pearson's correlation coefficient
- ICC (intraclass correlation coefficient)
- Kappa statistic

Example 4.1: Suppose we administer a language proficiency test and retest to a random sample of 10 students. Their scores from both time periods are shown below in columns B and C.

	A	B	C	D	E	F
1	Student	Test	Retest		Pearson's correlation coefficient:	
2	1	94	96		=CORREL(B2:B11,C2:C11)	
3	2	92	90		0.862349658743044	
4	3	89	91			
5	4	87	86			
6	5	87	88			
7	6	86	86			
8	7	87	89			
9	8	94	91			
10	9	85	84			
11	10	84	86			



One way to assess test-retest reliability is to compute **Pearson's correlation coefficient** between the two sets of scores. If the test is reliable and if none of the subjects have changed from Time 1 to Time 2 with regard to what is being measured, we should see a high correlation coefficient.

Questions:

1. What is the Pearson correlation coefficient for the example given above?

2. Does this indicate that this test is "reliable"? Explain.

3. In addition to computing the correlation coefficient, one should also compute the mean and standard deviation of the scores at each time period. Why?

The Pearson correlation coefficient is an acceptable measure of reliability, but it has been argued that a better measure of test-retest reliability for continuous data is the **intraclass correlation coefficient (ICC)**. One reason the ICC is preferred is that Pearson's correlation coefficient has been shown to overestimate reliability for small sample sizes. Another advantage the ICC has is that it can be calculated even when you administer the test at more than two time periods.

There are several versions of the ICC, but one that is typically used in examples such as this is computed as follows:

$$\text{ICC} = \frac{\text{MS}_{\text{Subject}} - \text{MS}_{\text{Error}}}{\text{MS}_{\text{Subject}} + (k-1)\text{MS}_{\text{Error}}},$$

where k = the number of time periods, $\text{MS}_{\text{Subject}}$ = the between-subjects mean square, and MS_{Error} = the mean square due to error after fitting a repeated measures ANOVA.

Let's compute the ICC for the data in Example 4.1.

Data in JMP:

Student	Time	score
1	1	94
2	1	92
3	1	89
4	1	87
5	1	87
6	1	86
7	1	87
8	1	94
9	1	85
10	1	84
1	2	96
2	2	90
3	2	91
4	2	86
5	2	88
6	2	86
7	2	89
8	2	91
9	2	84
10	2	86

Fitting the Model in JMP:

Select **Analyze > Fit Model** and enter the following:

Output from JMP:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	10	213.00000	21.3000	12.1329
Error	9	15.80000	1.7556	Prob > F
C. Total	19	228.80000		0.0005*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Student	9	9	212.80000	13.4684	0.0003*
Time	1	1	0.20000	0.1139	0.7435

$$ICC = \frac{MS_{Subject} - MS_{Error}}{MS_{Subject} + (k-1)MS_{Error}} =$$

In the previous example, the data were considered on a continuous scale. Note that when the data are measured on a binary scale, Cohen's **kappa statistic** should be used to estimate test-retest reliability; for nominal data with more than two categories, one can use Fleiss's kappa statistic. Finally, when the data are ordinal, one should use the **weighted kappa**.

Example 4.2: Suppose 10 nursing students are asked on two different occasions if they plan to work with older adults when they graduate.

Student	Time 1	Time 2
1	No	No
2	No	No
3	No	Yes
4	Yes	Yes
5	Yes	Yes
6	Yes	Yes
7	No	No
8	Yes	Yes
9	No	Yes
10	No	No

Cohen's kappa statistic is computed by first organizing the data as follows:

	Time 2: Yes	Time 2: No
Time 1: Yes	4	0
Time 1: No	2	4

Cohen's kappa statistic is a function of the number of agreements observed minus the number of agreements we expect by chance.

	Yes	No	Total
Agreements Observed			
Agreements Expected by Chance			

$$\kappa = \frac{\#\text{ of agreements observed} - \#\text{ of agreements expected by chance}}{n - \#\text{ of agreements expected by chance}} =$$

To interpret the kappa coefficient, consider the following discussion from wikipedia.org.

> Another factor is the number of codes. As number of codes increases, kappas become higher. Based on a simulation study, Bakeman and colleagues concluded that for fallible observers, values for kappa were lower when codes were fewer. And, in agreement with Sim & Wrights's statement concerning prevalence, kappas were higher when codes were roughly equiprobable. Thus Bakeman et al. concluded that "no one value of kappa can be regarded as universally acceptable."[10]:357 They also provide a computer program that lets users compute values for kappa specifying number of codes, their probability, and observer accuracy. For example, given equiprobable codes and observers who are 85% accurate, value of kappa are 0.49, 0.60, 0.66, and 0.69 when number of codes is 2, 3, 5, and 10, respectively.
>
> Nonetheless, magnitude guidelines have appeared in the literature. Perhaps the first was Landis and Koch,[11] who characterized values < 0 as indicating no agreement and 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement. This set of guidelines is however by no means universally accepted; Landis and Koch supplied no evidence to support it, basing it instead on personal opinion. It has been noted that these guidelines may be more harmful than helpful.[12] Fleiss's[13]:218 equally arbitrary guidelines characterize kappas over 0.75 as excellent, 0.40 to 0.75 as fair to good, and below 0.40 as poor.

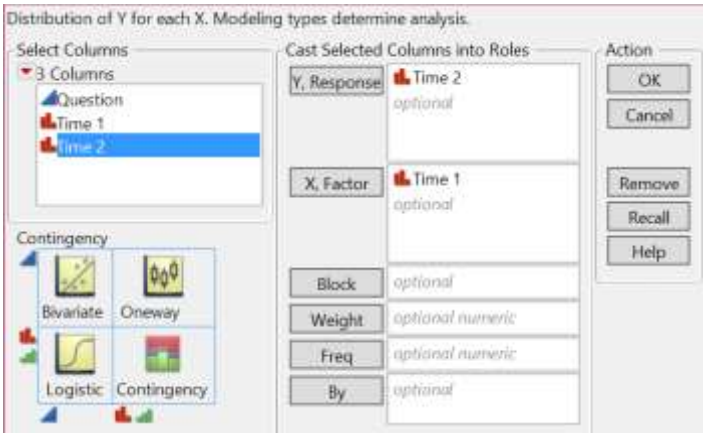
Question: Based on the kappa statistic obtained in Example 4.2, what can you say about the test-retest reliability of this question asked to nursing students?

Calculating the kappa statistic in JMP

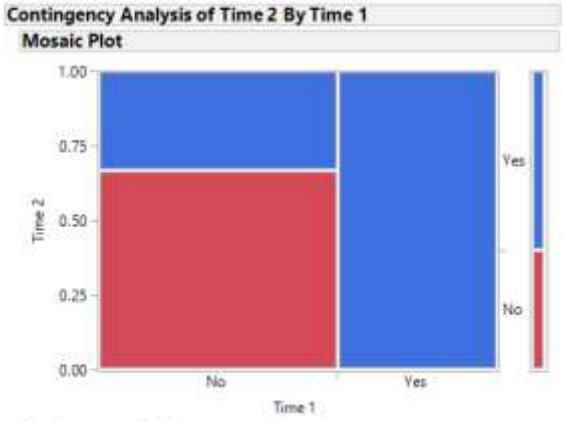
The data table is first arranged as follows:

Question	Time 1	Time 2
1	No	No
2	No	No
3	No	Yes
4	Yes	Yes
5	Yes	Yes
6	Yes	Yes
7	No	No
8	Yes	Yes
9	No	Yes
10	No	No

Select **Analyze > Fit Y by X** and enter the following:



JMP then displays the contingency table and mosaic plot:



Contingency Table

		Time 2		
	Count / Total % / Col % / Row %	No	Yes	Total
Time 1	No	4 40.00 100.00 66.67	2 20.00 33.33 33.33	6 60.00
	Yes	0 0.00 0.00 0.00	4 40.00 66.67 100.00	4 40.00
	Total	4 40.00	6 60.00	10

Tests

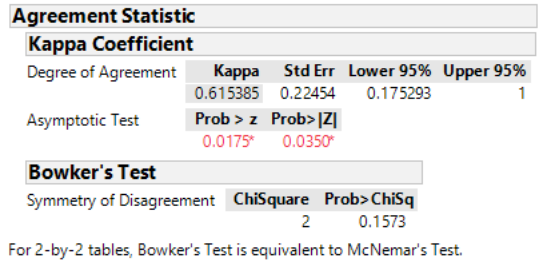
N	DF	-LogLike	RSquare (U)
10	1	2.9110317	0.4325

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.822	0.0158*
Pearson	4.444	0.0350*

Warning: Average cell count less than 5. LR ChiSquare suspect.

Fisher's Exact Test	Prob	Alternative Hypothesis
Left	1.0000	Prob(Time 2=Yes) is greater for Time 1=No than Yes
Right	0.0714	Prob(Time 2=Yes) is greater for Time 1=Yes than No
2-Tail	0.0762	Prob(Time 2=Yes) is different across Time 1

Finally, click on the red drop-down arrow that appears next to "Contingency Analysis of Time 2 by Time 1" and select **Agreement Statistic**. JMP returns the following.



Agreement Statistic

Kappa Coefficient

Degree of Agreement	Kappa	Std Err	Lower 95%	Upper 95%
	0.615385	0.22454	0.175293	1

Asymptotic Test	Prob > z	Prob>|Z|
	0.0175*	0.0350*

Bowker's Test

Symmetry of Disagreement	ChiSquare	Prob>ChiSq
	2	0.1573

For 2-by-2 tables, Bowker's Test is equivalent to McNemar's Test.

Note that if the data are nominal but not binary, these same steps in JMP will result in the calculation of Fleiss's kappa statistic.

Also, as mentioned earlier, if the data are ordinal, the **weighted kappa** should be used. This is a generalization of the kappa statistic that uses weights to quantify the relative difference between categories (essentially, disagreements that are further away from one another on the scale are weighted differently than disagreements that are closer together on the scale). The computation is not as easy as for the simple kappa statistic, so we won't discuss this in any more detail here. Note that several software packages exist (e.g., SAS, SPSS) that will compute both the kappa and weighted kappa statistics.

Parallel Forms Reliability

This involves creating a large set of questions that are believed to measure the same construct and then randomly dividing these questions into two sets (known as parallel forms). Both sets are then administered to the same group of people. The means, standard deviations, and correlations with other measures (when appropriate) should be compared to establish that the two forms are equivalent. The correlation between the two parallel forms can be used as the estimate of reliability (we want the scores on the two forms to be highly correlated).

Having parallel forms is useful for studies involving both a pre-test and post-test in which the researcher doesn't want to use the same form at both time periods (think of why they may want to avoid this). If parallel forms reliability can be established, then the researcher can use both forms in their study.

Question: If the researcher uses parallel forms, should they always use Form A for the pre-test and Form B for the post-test? Can you think of a different approach that might be better?

Internal Consistency Reliability

Earlier, we discussed methods for establishing test-retest reliability. Note that in some cases, it may be too expensive or time-consuming to administer a test twice. Or, one may not want "practice effects" to influence the results which is always a possibility when a measurement instrument is administered more than once.

In such cases, one is better off investigating **internal consistency reliability**. This involves administering a single measurement instrument (e.g., a survey questionnaire) to a group of people on only one occasion. Essentially, we judge the reliability of the instrument by estimating how well the items that measure the same construct yield similar results.

There are several different measures available for estimating internal consistency reliability. In particular, we will discuss the following in this handout:

- Average inter-item correlation
- Split-half reliability
- Cronbach's alpha

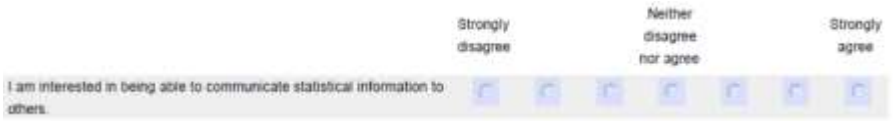
Sometimes, these measures are computed based on all items measured by the instrument; other times, these are used to establish the reliability associated with various constructs that are measured by different items within the same instrument. Next, we will introduce these measures of reliability in the context of an example.

Example 4.3. The Survey of Attitudes toward Statistics (SATS) measures six different components of students' attitudes about statistics. The survey overall consists of 36 questions. One of the constructs that the survey measures is students' "interest" in statistics, which is measured with the following four questions.

Interest – students' level of individual interest in statistics (4 items, new component):

12. I am interested in being able to communicate statistical information to others.
20. I am interested in using statistics.
23. I am interested in understanding statistical information.
29. I am interested in learning statistics.

All of the questions are measured on a 7-point scale, as shown below for only one question.



To score this survey, the data is first recoded so that 1 = strongly disagree, 4 = neither disagree nor agree, and 7 = strongly agree, etc. Then, the score for the "interest" component is computed by averaging the responses of these four questions. Other components are scored in the same way.

Calculating the Average Inter-Item Correlation

First, we will consider establishing reliability for only the "interest" scale using the average inter-item correlation to measure reliability.

These calculations can be carried out in Excel, as shown below.

	A	B	C	D	E	F
2						
3		**Subject ID**	**Q12**	**Q20**	**Q23**	**Q29**
4		1	5	5	4	3
5		2	2	1	1	1
6		3	4	6	5	6
7		4	4	6	5	6
8		5	6	5	6	6
9		6	1	5	5	5
10		7	2	2	4	4
11		8	5	5	6	5
12		9	2	3	3	3
13		10	1	3	2	5
14		11	5	5	5	5
15		12	4	3	5	5
16		13	6	5	6	6
17		14	3	5	6	6
18		15	5	5	5	5
19		16	4	4	4	4
20		17	3	3	5	1
21		18	4	5	6	7
22		19	7	6	7	7
23		20	4	5	5	5
24		21	6	6	6	6
25		22	4	5	4	5
26		23	4	5	4	3
27		24	4	4	5	4
28		25	2	4	4	5
29						
30		r 12, 20 :	=CORREL(C4:C28,D4:D28)	0.634666105026736		
31		r 12, 23 :	=CORREL(C4:C28,E4:E28)	0.688229269659616		
32		r 12, 29 :	=CORREL(C4:C28,F4:F28)	0.43720484789966		
33		r 20, 23:	=CORREL(D4:D28,E4:E28)	0.723510517142838		
34		r 20, 29 :	=CORREL(D4:D28,F4:F28)	0.699889631345202		
35		r 23, 29 :	=CORREL(E4:E28,F4:F28)	0.675274337759729		
36		Average r:	=AVERAGE(C30:C35)	0.643129118138963		

Note that the average inter-item correlation is .643, with the individual correlations ranging from .44 to .72. What does this imply about the reliability of the "interest" construct?

Calculating the Split-Half Reliability

To compute this measure of reliability, we correlate scores on one random half of the items on a survey (or test) with the scores on the other random half. Consider the calculation of this measure using the SATS data from Example 4.3 as shown below in Excel.

Excel sheet showing formulas:

H	I
Half 1: Q12, Q20 vs. Half 2: Q23, Q29	
Half 1	**Half 2**
=AVERAGE(C4:D4)	=AVERAGE(E4:F4)
=AVERAGE(C5:D5)	=AVERAGE(E5:F5)
=AVERAGE(C6:D6)	=AVERAGE(E6:F6)
=AVERAGE(C7:D7)	=AVERAGE(E7:F7)
=AVERAGE(C8:D8)	=AVERAGE(E8:F8)
=AVERAGE(C9:D9)	=AVERAGE(E9:F9)
=AVERAGE(C10:D10)	=AVERAGE(E10:F10)
=AVERAGE(C11:D11)	=AVERAGE(E11:F11)
=AVERAGE(C12:D12)	=AVERAGE(E12:F12)
=AVERAGE(C13:D13)	=AVERAGE(E13:F13)
=AVERAGE(C14:D14)	=AVERAGE(E14:F14)
=AVERAGE(C15:D15)	=AVERAGE(E15:F15)
=AVERAGE(C16:D16)	=AVERAGE(E16:F16)
=AVERAGE(C17:D17)	=AVERAGE(E17:F17)
=AVERAGE(C18:D18)	=AVERAGE(E18:F18)
=AVERAGE(C19:D19)	=AVERAGE(E19:F19)
=AVERAGE(C20:D20)	=AVERAGE(E20:F20)
=AVERAGE(C21:D21)	=AVERAGE(E21:F21)
=AVERAGE(C22:D22)	=AVERAGE(E22:F22)
=AVERAGE(C23:D23)	=AVERAGE(E23:F23)
=AVERAGE(C24:D24)	=AVERAGE(E24:F24)
=AVERAGE(C25:D25)	=AVERAGE(E25:F25)
=AVERAGE(C26:D26)	=AVERAGE(E26:F26)
=AVERAGE(C27:D27)	=AVERAGE(E27:F27)
=AVERAGE(C28:D28)	=AVERAGE(E28:F28)
correlation =	=CORREL(H4:H28,I4:I28)

Excel sheet showing values:

H	I
Half 1: Q12, Q20 vs. Half 2: Q23, Q29	
Half 1	**Half 2**
5	3.5
1.5	1
5	5.5
5	5.5
5.5	6
3	5
2	4
5	5.5
2.5	3
2	3.5
5	5
3.5	5
5.5	6
4	6
5	5
4	4
3	3
4.5	6.5
6.5	7
4.5	5
6	6
4.5	4.5
4.5	3.5
4	4.5
3	4.5
correlation =	0.752430852

In this example, the split-half correlation is $r_{split\text{-}half} = .75$. One problem with the split-half method is that reducing the number of items on a survey (or test) generally reduces the reliability. Note that each of our split-half versions has only half the items that the full test has for measuring "interest." To correct for this, you should apply the Spearman-Brown correction:

$$r_{SB} = \frac{2 \times r_{split-half}}{1 + r_{split-half}} =$$

Calculating Cronbach's Alpha

One problem with the split-half method is that depending on which "split-halves" are used, the reliability estimate will change. One solution is to compute the Spearman-Brown corrected split-half reliability coefficients for each of the possible split-halves and then find their mean. This can be computed in Excel, as is shown below for the SATS data from Example 4.3:

Excel sheet showing formulas:

Excel sheet showing values:

Half 1: Q12, Q20 vs. Half 2: Q23, Q29	
Half 1	Half 2
5	3.5
1.5	1
5	5.5
5	5.5
5.5	6
3	5
2	4
5	5.5
2.5	3
2	3.5
5	5
3.5	5
5.5	6
4	6
5	5
4	4
3	3
4.5	6.5
6.5	7
4.5	5
6	6
4.5	4.5
4.5	3.5
4	4.5
3	4.5
correlation =	0.752430852
with SB correction:	0.85872815

Half 1: Q12, Q23 vs. Half 2: Q20, Q29	
Half 1	Half 2
4.5	4
1.5	1
4.5	6
4.5	6
6	5.5
3	5
3	3
5.5	5
2.5	3
1.5	4
5	5
4.5	4
6	5.5
4.5	5.5
5	5
4	4
4	2
5	6
7	6.5
4.5	5
6	6
4	5
4	4
4.5	4
3	4.5
correlation =	0.712579161
with SB correction:	0.832170771

Half 1: Q12, Q29 vs. Half 2: Q20, Q23	
Half 1	Half 2
4	4.5
1.5	1
5	5.5
5	5.5
6	5.5
3	5
3	3
5	5.5
2.5	3
3	2.5
5	5
4.5	4
6	5.5
4.5	5.5
5	5
4	4
2	4
5.5	5.5
7	6.5
4.5	5
6	6
4.5	4.5
3.5	4.5
4	4.5
3.5	4
correlation =	0.857169758
with SB correction:	0.923092522

Cronbach's alpha can be thought of as the average of all possible split-half estimates (note that these two are equivalent only when the item standard deviations are equal).

Though the computations above provide insight into what Cronbach's alpha is measuring, it is never calculated this way in practice. Instead, the following formulas are typically used:

Formula #1: $$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_X^2}\right)$$

where K is the number of components (items), σ_X^2 is the variance of the observed total scores, and $\sigma_{Y_i}^2$ is the variance of the scores on the i^{th} item.

This formula is used in Excel as follows:

	A	B	C	D	E	F	G	H	I
1			Item 1	Item 2	Item 3	Item 4		Total Test Scores:	
2		Subject 1	5	5	4	3		=SUM(C2:F2)	
3		Subject 2	2	1	1	1		=SUM(C3:F3)	
4		Subject 3	4	6	5	6		=SUM(C4:F4)	
5		Subject 4	4	6	5	6		=SUM(C5:F5)	
6		Subject 5	6	5	6	6		=SUM(C6:F6)	
7		Subject 6	1	5	5	5		=SUM(C7:F7)	
8		Subject 7	2	2	4	4		=SUM(C8:F8)	
9		Subject 8	5	5	6	5		=SUM(C9:F9)	
10		Subject 9	2	3	3	3		=SUM(C10:F10)	
11		Subject 10	1	3	2	5		=SUM(C11:F11)	
12		Subject 11	5	5	5	5		=SUM(C12:F12)	
13		Subject 12	4	3	5	5		=SUM(C13:F13)	
14		Subject 13	6	5	6	6		=SUM(C14:F14)	
15		Subject 14	3	5	6	6		=SUM(C15:F15)	
16		Subject 15	5	5	5	5		=SUM(C16:F16)	
17		Subject 16	4	4	4	4		=SUM(C17:F17)	
18		Subject 17	3	3	5	1		=SUM(C18:F18)	
19		Subject 18	4	5	6	7		=SUM(C19:F19)	
20		Subject 19	7	6	7	7		=SUM(C20:F20)	
21		Subject 20	4	5	5	5		=SUM(C21:F21)	
22		Subject 21	6	6	6	6		=SUM(C22:F22)	
23		Subject 22	4	5	4	5		=SUM(C23:F23)	
24		Subject 23	4	5	4	3		=SUM(C24:F24)	
25		Subject 24	4	4	5	4		=SUM(C25:F25)	
26		Subject 25	2	4	4	5		=SUM(C26:F26)	
27									
28		Item variances:	=VAR(C2:C26)	=VAR(D2:D26)	=VAR(E2:E26)	=VAR(F2:F26)		Variance of total test scores:	=VAR(H2:H26)
29									
30		k:	4						
31		Cronbach's alpha:	=C30/(C30-1)*(1-SUM(C28:F28)/I28)						

Excel sheet showing values:

	A	B	C	D	E	F	G	H	I
1			Item 1	Item 2	Item 3	Item 4		**Total Test Scores:**	
2		Subject 1	5	5	4	3		17	
3		Subject 2	2	1	1	1		5	
4		Subject 3	4	6	5	6		21	
5		Subject 4	4	6	5	6		21	
6		Subject 5	6	5	6	6		23	
7		Subject 6	1	5	5	5		16	
8		Subject 7	2	2	4	4		12	
9		Subject 8	5	5	6	5		21	
10		Subject 9	2	3	3	3		11	
11		Subject 10	1	3	2	5		11	
12		Subject 11	5	5	5	5		20	
13		Subject 12	4	3	5	5		17	
14		Subject 13	6	5	6	6		23	
15		Subject 14	3	5	6	6		20	
16		Subject 15	5	5	5	5		20	
17		Subject 16	4	4	4	4		16	
18		Subject 17	3	3	5	1		12	
19		Subject 18	4	5	6	7		22	
20		Subject 19	7	6	7	7		27	
21		Subject 20	4	5	5	5		19	
22		Subject 21	6	6	6	6		24	
23		Subject 22	4	5	4	5		18	
24		Subject 23	4	5	4	3		16	
25		Subject 24	4	4	5	4		17	
26		Subject 25	2	4	4	5		15	
27									
28		**Item variance**	2.526666667	1.67333	1.79333	2.46		**Variance of total test scores:**	24.35666667
29									
30		**k:**	4						
31		**Cronbach's alpha:**	0.870580722						

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_X^2}\right) =$$

Formula #2:

$$\alpha = \frac{K\overline{c}}{\left(\overline{v} + \left(K-1\right)\overline{c}\right)}$$

where K is the number of components (items), $\overline{v}$ is the average variance of each item, and $\overline{c}$ is the average of all covariances between the items.

This formula is used in Excel as follows:

	A	B	C	D	E	F
1			Item 1	Item 2	Item 3	Item 4
2		Subject 1	5	5	4	3
3		Subject 2	2	1	1	1
4		Subject 3	4	6	5	6
5		Subject 4	4	6	5	6
6		Subject 5	6	5	6	6
7		Subject 6	1	5	5	5
8		Subject 7	2	2	4	4
9		Subject 8	5	5	6	5
10		Subject 9	2	3	3	3
11		Subject 10	1	3	2	5
12		Subject 11	5	5	5	5
13		Subject 12	4	3	5	5
14		Subject 13	6	5	6	6
15		Subject 14	3	5	6	6
16		Subject 15	5	5	5	5
17		Subject 16	4	4	4	4
18		Subject 17	3	3	5	1
19		Subject 18	4	5	6	7
20		Subject 19	7	6	7	7
21		Subject 20	4	5	5	5
22		Subject 21	6	6	6	6
23		Subject 22	4	5	4	5
24		Subject 23	4	5	4	3
25		Subject 24	4	4	5	4
26		Subject 25	2	4	4	5
27						
28		**Item variances:**	=VAR(C2:C26)	=VAR(D2:D26)	=VAR(E2:E26)	=VAR(F2:F26)
29						
30		**k:**	4			
31		**Cronbach's alpha:**	=C30/(C30-1)*(1-SUM(C28:F28)/I28)			
32						
33						
34						
35		**Covariance Matrix:**				
36			Item 1	Item 2	Item 3	Item 4
37		Item 1	=COVARIANCE.S(C2:C26,C2:C26)	=COVARIANCE.S(C2:C26,D2:D26)	=COVARIANCE.S(C2:C26,E2:E26)	=COVARIANCE.S(C2:C26,F2:F26)
38		Item 2	=COVARIANCE.S(C2:C26,D2:D26)	=COVARIANCE.S(D2:D26,D2:D26)	=COVARIANCE.S(D2:D26,E2:E26)	=COVARIANCE.S(D2:D26,F2:F26)
39		Item 3	=COVARIANCE.S(E2:E26,C2:C26)	=COVARIANCE.S(D2:D26,E2:E26)	=COVARIANCE.S(E2:E26,E2:E26)	=COVARIANCE.S(E2:E26,F2:F26)
40		Item 4	=COVARIANCE.S(C2:C26,F2:F26)	=COVARIANCE.S(D2:D26,F2:F26)	=COVARIANCE.S(E2:E26,F2:F26)	=COVARIANCE.S(F2:F26,F2:F26)
41						
42		**Average covariance:**	=AVERAGE(D37,E37,F37, E38,F38,F39)			
43		**Average variance:**	=AVERAGE(C28:F28)			
44		**Cronbach's alpha:**	=C30*C43/(C43+(C30-1)*C42)			

Excel sheet showing values:

	A	B	C	D	E	F
1			Item 1	Item 2	Item 3	Item 4
2		Subject 1	5	5	4	3
3		Subject 2	2	1	1	1
4		Subject 3	4	6	5	6
5		Subject 4	4	6	5	6
6		Subject 5	6	5	6	6
7		Subject 6	1	5	5	5
8		Subject 7	2	2	4	4
9		Subject 8	5	5	6	5
10		Subject 9	2	3	3	3
11		Subject 10	1	3	2	5
12		Subject 11	5	5	5	5
13		Subject 12	4	3	5	5
14		Subject 13	6	5	6	6
15		Subject 14	3	5	6	6
16		Subject 15	5	5	5	5
17		Subject 16	4	4	4	4
18		Subject 17	3	3	5	1
19		Subject 18	4	5	6	7
20		Subject 19	7	6	7	7
21		Subject 20	4	5	5	5
22		Subject 21	6	6	6	6
23		Subject 22	4	5	4	5
24		Subject 23	4	5	4	3
25		Subject 24	4	4	5	4
26		Subject 25	2	4	4	5
27						
28		**Item varianc**	2.526666667	1.673333333	1.793333333	2.46
29						
30		**k:**	4			
31		**Cronbach's alpha:**	0.870580722			
32						
33						
34						
35		**Covariance Matrix:**				
36			Item 1	Item 2	Item 3	Item 4
37		Item 1	2.526666667	1.305	1.465	1.09
38		Item 2	1.305	1.673333333	1.253333333	1.42
39		Item 3	1.465	1.253333333	1.793333333	1.418333333
40		Item 4	1.09	1.42	1.418333333	2.46
41						
42		**Average covariance:**	1.325277778			
43		**Average variance:**	2.113333333			
44		**Cronbach's**	0.870580722			

$$\alpha = \frac{K\bar{c}}{\left(\bar{v} + (K-1)\bar{c}\right)}$$

Formula #3: $\alpha_{\text{standardized}} == \dfrac{K\bar{r}}{(1+(K-1)\bar{r})}$

where K is the number of components (items), $\bar{v}$ is the average variance of each item, and $\bar{r}$ is the mean of the correlation coefficients.

This is known as the **Standardized Cronbach's alpha** and should be used if the items are measured on different scales.

This can also be computed in Excel:

	A	B	C	D	E	F
1			Item 1	Item 2	Item 3	Item 4
2		Subject 1	5	5	4	3
3		Subject 2	2	1	1	1
4		Subject 3	4	6	5	6
5		Subject 4	4	6	5	6
6		Subject 5	6	5	6	6
7		Subject 6	1	5	5	5
8		Subject 7	2	2	4	4
9		Subject 8	5	5	6	5
10		Subject 9	2	3	3	3
11		Subject 10	1	3	2	5
12		Subject 11	5	5	5	5
13		Subject 12	4	3	5	5
14		Subject 13	6	5	6	6
15		Subject 14	3	5	6	6
16		Subject 15	5	5	5	5
17		Subject 16	4	4	4	4
18		Subject 17	3	3	5	1
19		Subject 18	4	5	6	7
20		Subject 19	7	6	7	7
21		Subject 20	4	5	5	5
22		Subject 21	6	6	6	6
23		Subject 22	4	5	4	5
24		Subject 23	4	5	4	3
25		Subject 24	4	4	5	4
26		Subject 25	2	4	4	5
27						
28		Item variances:	=VAR(C2:C26)	=VAR(D2:D26)	=VAR(E2:E26)	=VAR(F2:F26)
29						
30		k:	4			
31		Cronbach's alpha:	=C30/(C30-1)*(1-SUM(C28:F28)/I28)			
32						
33						
34						
35	Correlation Matrix:					
36		Item 1	Item 2	Item 3	Item 4	
37	Item 1	=CORREL(C2:C26,C2:C26)	=CORREL(C2:C26,D2:D26)	=CORREL(C2:C26,E2:E26)	=CORREL(C2:C26,F2:F26)	
38	Item 2	=CORREL(C2:C26,D2:D26)	=CORREL(D2:D26,D2:D26)	=CORREL(D2:D26,E2:E26)	=CORREL(D2:D26,F2:F26)	
39	Item 3	=CORREL(C2:C26,E2:E26)	=CORREL(D2:D26,E2:E26)	=CORREL(E2:E26,E2:E26)	=CORREL(E2:E26,F2:F26)	
40	Item 4	=CORREL(C2:C26,F2:F26)	=CORREL(D2:D26,F2:F26)	=CORREL(E2:E26,F2:F26)	=CORREL(F2:F26,F2:F26)	
41						
42	Average correlation:	=AVERAGE(C37,D37,E37,D3				
43	Cronbach's alpha:	=C30*B42/(1+(C30-1)*B42)				

Excel sheet showing values:

	A	B	C	D	E	F
1			Item 1	Item 2	Item 3	Item 4
2		Subject 1	5	5	4	3
3		Subject 2	2	1	1	1
4		Subject 3	4	6	5	6
5		Subject 4	4	6	5	6
6		Subject 5	6	5	6	6
7		Subject 6	1	5	5	5
8		Subject 7	2	2	4	4
9		Subject 8	5	5	6	5
10		Subject 9	2	3	3	3
11		Subject 10	1	3	2	5
12		Subject 11	5	5	5	5
13		Subject 12	4	3	5	5
14		Subject 13	6	5	6	6
15		Subject 14	3	5	6	6
16		Subject 15	5	5	5	5
17		Subject 16	4	4	4	4
18		Subject 17	3	3	5	1
19		Subject 18	4	5	6	7
20		Subject 19	7	6	7	7
21		Subject 20	4	5	5	5
22		Subject 21	6	6	6	6
23		Subject 22	4	5	4	5
24		Subject 23	4	5	4	3
25		Subject 24	4	4	5	4
26		Subject 25	2	4	4	5
27						
28		**Item varianc**	2.526666667	1.673333333	1.793333333	2.46
29						
30		**k:**	4			
31		**Cronbach's alpha:**	0.870580722			
32						
33						
34						
35	**Correlation Matrix:**					
36		Item 1	Item 2	Item 3	Item 4	
37	Item 1	1	0.634666105	0.68822927	0.437204848	
38	Item 2	0.634666105	1	0.723510517	0.699889631	
39	Item 3	0.68822927	0.723510517	1	0.675274338	
40	Item 4	0.437204848	0.699889631	0.675274338	1	
41						
42	**Average correlation:**	0.643129118				
43	**Cronbach's alpha:**	0.878175591				

$$\alpha_{\text{standardized}} == \frac{K\bar{r}}{(1+(K-1)\bar{r})} =$$

Interpreting Cronbach's Alpha

To interpret Cronbach's alpha, consider the following discussion from wikipedia.org.

A commonly accepted[citation needed] rule of thumb for describing internal consistency using Cronbach's alpha is as follows,[16][17] however, a greater number of items in the test can artificially inflate the value of alpha[8] and a sample with a narrow range can deflate it, so this rule of thumb should be used with caution:

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent (High-Stakes testing)
$0.7 \leq \alpha < 0.9$	Good (Low-Stakes testing)
$0.6 \leq \alpha < 0.7$	Acceptable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

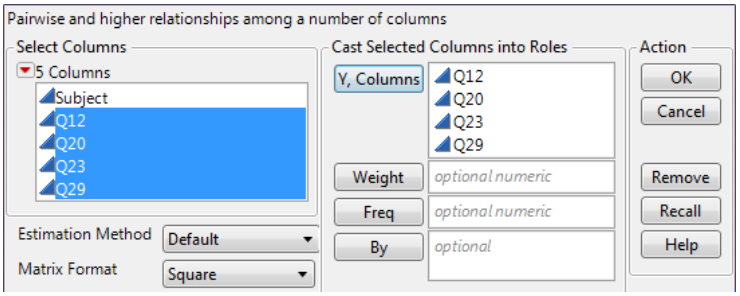
- There are some problems with this somewhat arbitrary rule of thumb. For example, Cronbach's alpha tends to increase with the number of items on the scale; so, a high alpha doesn't necessarily that the measure is "reliable."
- Cronbach's alpha is *not* a test for unidimensionality.

Computing Cronbach's Alpha with Software

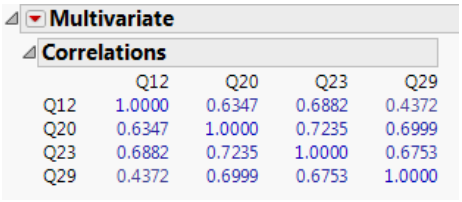
Software packages such as SAS, SPSS, or JMP can also be used to compute Cronbach's alpha. To calculate this in JMP, enter the data as follows:

Subject	Q12	Q20	Q23	Q29
1	5	5	4	3
2	2	1	1	1
3	4	6	5	6
4	4	6	5	6
5	6	5	6	6
6	1	5	5	5
7	2	2	4	4
8	5	5	6	5
9	2	3	3	3
10	1	3	2	5
11	5	5	5	5
12	4	3	5	5
13	6	5	6	6
14	3	5	6	6
15	5	5	5	5
16	4	4	4	4
17	3	3	5	1
18	4	5	6	7
19	7	6	7	7
20	4	5	5	5
21	6	6	6	6
22	4	5	4	5
23	4	5	4	3
24	4	4	5	4
25	2	4	4	5

Select **Analyze > Multivariate Methods > Multivariate**. Then, enter the following:



JMP returns the correlation matrix:

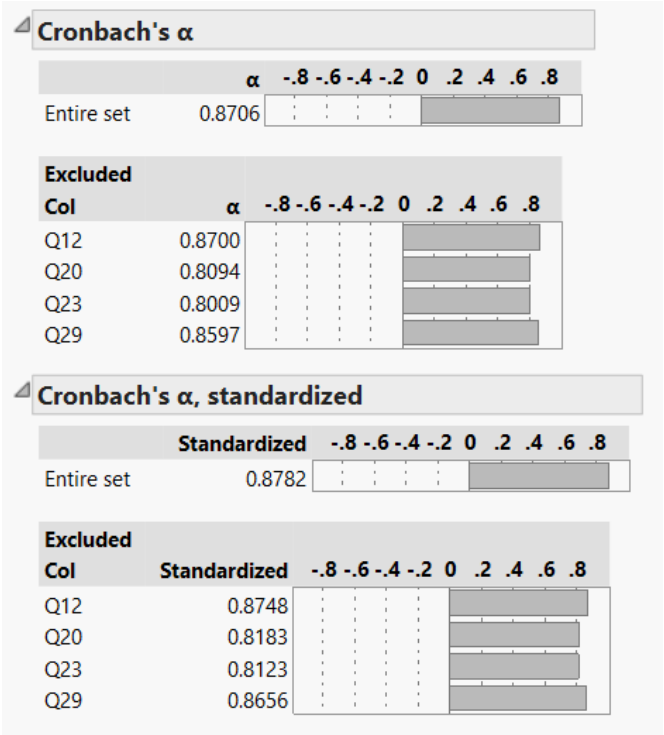


Multivariate

Correlations

	Q12	Q20	Q23	Q29
Q12	1.0000	0.6347	0.6882	0.4372
Q20	0.6347	1.0000	0.7235	0.6999
Q23	0.6882	0.7235	1.0000	0.6753
Q29	0.4372	0.6999	0.6753	1.0000

From the red drop-down arrow next to **Multivariate**, select **Item Reliability** and then either **Cronbach's alpha** or **Standardized alpha.** The JMP output for the SATS data from Example 4.3 is shown below.



Cronbach's α

	α
Entire set	0.8706

Excluded Col	α
Q12	0.8700
Q20	0.8094
Q23	0.8009
Q29	0.8597

Cronbach's α, standardized

	Standardized
Entire set	0.8782

Excluded Col	Standardized
Q12	0.8748
Q20	0.8183
Q23	0.8123
Q29	0.8656

Some Words of Caution

Note that Cronbach's alpha is affected by reverse worded items. For example, consider another component measured in the SATS discussed in Example 4.3. This survey also measures students' perceived "value" of statistics using the following questions.

Value – students' attitudes about the usefulness, relevance, and worth of statistics in personal and professional life (9 items; .74 to .90)):

7.* Statistics is worthless.
9. Statistics should be a required part of my professional training.
10. Statistical skills will make me more employable.
13.* Statistics is not useful to the typical professional.
16.* Statistical thinking is not applicable in my life outside my job.
17. I use statistics in my everyday life.
21.* Statistics conclusions are rarely presented in everyday life.
25.* I will have no application for statistics in my profession.
33.* Statistics is irrelevant in my life.

If you don't value statistics, you might answer Question 7 with "strongly agree" which is coded as a 7. On the other hand, you might answer Question 9 with "strongly disagree" which is coded as a 1. If you have reverse worded items, then you also have to reverse the way in which they are scored before you conduct a reliability analysis. The creators of the SATS include the following instructions on scoring their survey.

Component (subscale) scores on the SATS-36 are formed by

1. Reversing the responses to the negatively worded items indicated with an asterisk* (1 becomes 7, 2 becomes 6, etc.),
2. Summing the item responses within each component, and
3. Dividing by the number of items within each component.