


---

## 26. Reliability and Validity of Scales

---

When we looked at **inter-rater reliability** our focus was on differences in the raters and, to a lesser degree, whether the survey items differentiated between subjects. In this section we change focus to the research instrument itself. Our goal will be to see how well respondent answers “hang together.”

The goal of a survey is to measure a variable or variables that are relevant to your research objectives. Sometimes these variables are easy to **directly** measure—things like income, or blood pressure, or weight loss. Other times, though, variables are more subtle. This section is about using a **group** of survey questions, or **items**, to classify individuals according to psychological or social traits that can’t be directly measured.



When social scientists do this, they develop a set of questions that people in a particular group—say impulsive people or people sharing a particular socio-economic status—will answer in a similar way. Instead of measuring the variable **directly**, then, the measurement is **indirect**.

## 26.1. Example.


*The Human Resources Department at Mechanics R Us develops a set of six questions related to job satisfaction.*

	Agree Strongly	Agree	Disagree	Disagree Strongly
No one outside of work cares what I do here.				
I look forward to coming to work every day.				
Friday is the best day of the week.				
The most important thing about my job is that it pays				
At the end of the workday, I feel good about what I've				
People respect what I do at work.				

*The Director is interested in whether employees consistently answer these questions.*

**Solution.** In order to answer her questions, the HR Director randomly selects 20 employees and administers the questionnaire. Since some of the questions relate to positive views about work and some about negative views, she **scores** the answers so that **positive views** score higher.

	Agree Strongly	Agree	Disagree	Disagree Strongly
No one outside of work cares what I do here.	0	1	2	3
I look forward to coming to work every day.	3	2	1	0
Friday is the best day of the week.	0	1	2	3
The most important thing about my job is that it pays the bills.	0	1	2	3
At the end of the workday, I feel good about what I've accomplished.	3	2	1	0
People respect what I do at work.	3	2	1	0



The scores can then range from a low of zero to a high of eighteen.  
With this scoring, she obtains the following results.

Subject	Q1	Q2	Q3	Q4	Q5	Q6
1	0	0	1	1	0	0
2	2	3	2	3	2	2
3	0	1	1	2	2	1
4	0	0	2	1	1	1
5	2	3	1	2	3	2
6	3	3	2	0	1	3
7	0	0	1	2	3	1
8	2	3	2	2	3	3
9	3	3	2	2	2	2
10	2	3	2	1	2	3
11	3	2	2	3	1	2
12	3	1	3	3	3	0
13	0	0	3	2	0	0
14	3	3	3	3	2	3
15	0	0	1	1	1	2
16	2	0	1	0	0	1
17	0	1	1	0	3	2
18	3	0	3	2	1	3
19	3	2	3	2	2	2
20	3	2	1	2	2	3

Cronbach's alpha is a measure of the internal consistency of this scale. It addresses the question of whether or not respondents are giving consistent answers.

Cronbach's Alpha=												Interpretation	
0.705608099												Cronbach's Alpha	Internal Consistency
Number of Subjects	60											alpha >= 0.9	Excellent
Number of Questions	6											0.8 <= alpha < 0.9	Good
<p><i>Do not edit the RED cells. Put your data in the Green Cells. You may have up to 20 items and up to 1000 subjects. Leave BLANK columns for which you have no questions. EXCLUDE subjects who do not have complete responses (i.e., did not answer every question)</i></p>												0.7 <= alpha < 0.8	Acceptable
												0.6 <= alpha < 0.7	Questionable
												0.5 <= alpha < 0.6	Poor
												alpha < 0.5	Unacceptable
		Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10		
Item Mean		2.6	2.666666667	2.65	2.75	2.833333333	1.8						
Item StdDev		1.356465997	1.362187783	1.122868351	1.299038106	1.614173335	1.029563014						
Cronbach Alpha w/o this item		0.580655076	0.553811946	0.600484918	0.578961933	0.691063137	0.415061884						

In this case, we have an  $\alpha$  of 0.705, which tells us that the internal consistency is acceptable.

Another question of interest to the HR director is whether or not the questions distinguish between the subjects. After all, if they all reported




the same job satisfaction, it wouldn't be a useful scale even it were internally consistent!

To check whether or not the scale differentiates between the employees, we could do an ANOVA, transposing the data so that the employee responses become the columns. If we do this, the resulting  $p$ -value is less than 0.01%, so we can be quite confident that the scale differentiates between employees. See the example spreadsheet for this section, which uses the data analysis tool to do the ANOVA.


When we use a scale, we say that the items on the instrument measure a **latent variable**.

As with any measurement, the broad goals in capturing latent variables are:



- 
- *Standardization*—all steps in data collection are standardized and consistent;
  - *Objectivity*—data gathering minimizes subjective biases from the observed and the observer;
  - *Test normalization*—test results from a large group provides the basis for comparison with individual results;
  - *Reliability*—multiple tests give the same conclusions;
  - *Validity*—the data actually measure the intended variable.

For example, suppose your variable is "introversion." You can't just ask someone if they are introverted, since that requires a subjective opinion on the part of the respondent. Instead, you might ask a *series of questions* designed to capture the concept of introversion. In this case, introversion is be a *latent variable* which is captured by a pattern of answers to the questions. Thus, in this case the individual questions, or items on the survey, don't represent variables at all. Instead, taken together, they collectively classify the extent to which individuals exhibit the latent variable. The set of questions used to capture the latent variable are sometimes called a *scale* since they are used to measure the degree to which the subject shares the trait with other people. Early scales for introversion asked questions like the following:

- 
- Do you suddenly feel shy when you want to talk to an attractive stranger?
  - Generally do you prefer reading to meeting people?
  - Do you prefer to have few but special friends?
  - Do you find it hard to really enjoy yourself at a lively party?
  - Do you like talking to people so much that you never miss a chance of talking to a stranger?
  - Can you easily get some life into a dull party?
  - Do you look forward to speaking in public?

You might expect an introvert, for example, to answer "yes" to the first four questions and "no" to the last three, so you would score the scale accordingly, giving a score of "1" to a "yes" on the first four questions and a score of "1" to a "no" on the last three. This set of questions thus appears to be **valid** since they all seem to deal with a particular social anxiety experienced by introverts. If you were testing for "extroversion,"

you'd just reverse the scoring, of course.

One of the ideas with writing scales for latent variables is to repeatedly ask similar questions phrased in different ways, some positive and some negative. Sometimes scales will include other questions as distractors, or even other questions looking for other variables.

Sometimes the latent variable turns out to consist of several less obvious sub-variables or *dimensions*. As an example, Sir Lawrence Olivier was well-known to be painfully shy, a characteristic shared by many introverts. Yet he clearly had no problem speaking in public—that was his career! This famous example illustrates that "communication anxiety" might be one dimension of introversion and "shyness" another. Indeed, it's not all uncommon for latent variables to have more than one dimension.

It's possible to analyze scales for dimensionality using something called factor analysis, but that's beyond the scope of this course.

It's also possible to test for validity. For example, simple inspection of our list of questions above verifies that it includes elements of intro-

verted/extroverted behavior—**content validity**.

Another kind of validity, **construct validity**, involves how well our scale actually measures the variable. A simple way to test for this might involve comparing our proposed scale with another, different scale for the *same trait*. A high correlation would increase our confidence in the validity of both scales.

Another way to test for construct validity might be to see how well responses to our scale correlate with responses to a scale designed to measure a *different* latent variable, say impulsiveness. Presumably, these are *different* variables, so we'd expect the correlation to be low: there's no relation between intro/extroversion and impulsiveness. A high correlation with a scale measuring an ostensibly different trait would thus call into question the validity of our proposed scale.

Finally, our scale should have **predictive** ability, or **criterion validity**. We should be able to use our scale to predict outcomes. A person scoring higher on an extroversion scale, for example, should seek out social situations that provide opportunities for lots of interaction and meeting

new people.

**Cronbach's alpha** is a particular test for the *reliability* of a scale. It is similar to the analysis of variance in that it compares the variability within the items to the overall variability of the entire scale. Generally speaking, the higher the value of alpha, the more reliable the scale. The generally accepted practice is

$\alpha \geq 0.9$	excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

For reference, the formula for Cronbach's alpha is:

$$\alpha = \left( \frac{k}{k-1} \right) \left( 1 - \frac{\sum_{i=1}^k \sigma_{y_i}^2}{\sigma_x^2} \right)$$

where  $k$  is the number of items on the scale,  $\sigma_{y_i}^2$  are the within-item variances, and  $\sigma_x^2$  is the total variance.

Cronbach's alpha can be computationally complex, but spreadsheets or statistical programs make it relatively easy to calculate. The [AnalyzeThis](#) spreadsheet included in the course materials has a tab for using Cronbach's alpha to analyze data from a scale. The spreadsheet even includes a calculation of alpha where each item in turn is omitted from the scale. If the alpha is unchanged by omitting an item, it might be redundant and a candidate to remove from the scale.

## 26.2. Example.

*A sample of 60 students take the following survey.*

Please circle the answer that best describes your answer.

1. Statistics are dull.
  1. Strongly agree
  2. Agree
  3. Neutral
  4. Disagree
  5. Strongly disagree
2. Statistics are useful.
  1. Strongly agree
  2. Agree
  3. Neutral
  4. Disagree
  5. Strongly disagree
3. Statistics are important to understand.
  1. Strongly agree
  2. Agree
  3. Neutral
  4. Disagree
  5. Strongly disagree
4. People sometimes lie with statistics.
  1. Strongly agree
  2. Agree
  3. Neutral
  4. Disagree
  5. Strongly disagree
5. I love working with numbers.
  1. Strongly agree
  2. Agree
  3. Neutral
  4. Disagree
  5. Strongly disagree


*The results of this survey are stored in the spreadsheet **AnalyzeThis** on the tab **Cronbach**. Do the questions at left appear to all be dealing with the same latent variable? How reliable is this questionnaire? Suggest at least one way to improve the reliability.*



B1     $= (B4 / (B4 - 1)) * (1 - (B17 / B19))$

	A	B	C	D	E	F	G	H	I	J	K
1	<b>Cronbach's Alpha=</b>	<b>0.689371957</b>								<b>Interpretation</b>	
2										<b>Cronbach's Alpha</b>	<b>Internal Consistency</b>
3	<b>Number of Subjects</b>	<b>60</b>								alpha >= 0.9	Excellent
4	<b>Number of Questions</b>	<b>5</b>								0.8 <= alpha < 0.9	Good
5										0.7 <= alpha < 0.8	Acceptable
6										0.6 <= alpha < 0.7	Questionable
7										0.5 <= alpha < 0.6	Poor
8										alpha < 0.5	Unacceptable
6	<p><i>Do not edit the RED cells.  Put your data in the Green Cells.  You may have up to 20 items and up to 1000 subjects.  Leave BLANK columns for which you have no questions.  EXCLUDE subjects who do not have complete responses (i.e., did not answer every question)</i></p>										
12		<b>Question 1</b>	<b>Question 2</b>	<b>Question 3</b>	<b>Question 4</b>	<b>Question 5</b>	<b>Question 6</b>	<b>Question 7</b>	<b>Question 8</b>	<b>Question 9</b>	<b>Question 10</b>
13	<b>Item Mean</b>	3.133333333	3.266666667	3.016666667	3.283333333	3.4					
15	<b>Item StdDev</b>	1.071862346	0.997775303	1.072251007	1.126819516	1.704894914					
24	<b>Cronbach Alpha w/o this item</b>	0.618953448	0.58079096	0.577007769	0.561825518	0.839004724					
26		<b>Question 1</b>	<b>Question 2</b>	<b>Question 3</b>	<b>Question 4</b>	<b>Question 5</b>	<b>Question 6</b>	<b>Question 7</b>	<b>Question 8</b>	<b>Question 9</b>	<b>Question 10</b>
27	<b>Subject1</b>	4	3	4	4	1					
28	<b>Subject2</b>	4	3	3	3	5					
29	<b>Subject3</b>	5	5	5	4	5					
30	<b>Subject4</b>	3	4	4	5	5					
31	<b>Subject5</b>	3	4	3	3	1					
32	<b>Subject6</b>	3	3	2	3	5					
33	<b>Subject7</b>	2	3	2	2	5					
34	<b>Subject8</b>	4	5	4	5	1					
35	<b>Subject9</b>	4	5	3	5	5					
36	<b>Subject10</b>	5	3	3	5	5					
37	<b>Subject11</b>	2	2	1	1	1					

Ready    Chi-square, 1-way    Chi-square, 2-way    **Cronbach**    +



Cronbach's alpha is not a test for validity. It is also not a test for dimensionality. It only provides guidelines for the reliability of the scale.