

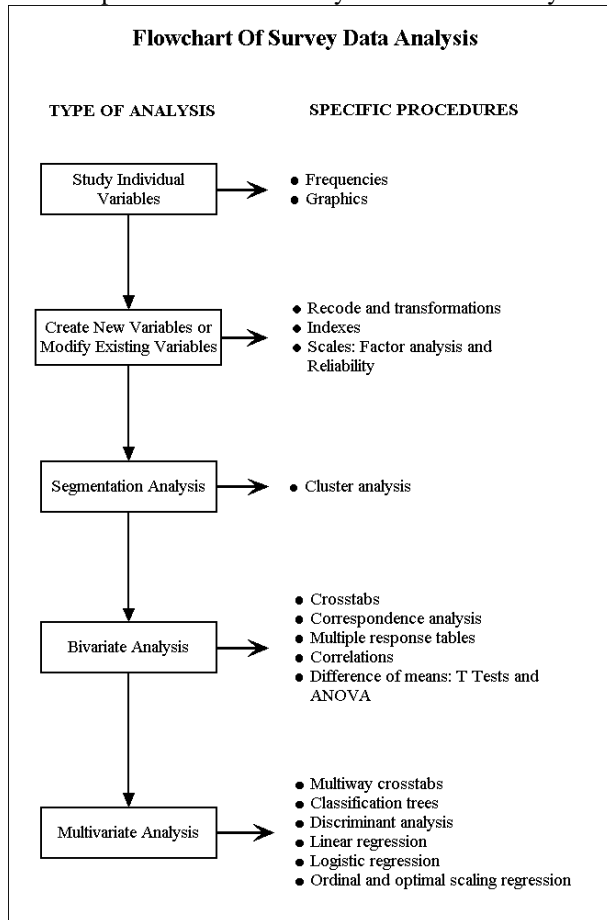
# Surveys- New Variables, Reliability and Validity and Factor Analysis -Using IBM SPSS

## Objective

Discuss the concepts of reliability and validity. Create an index variable. Use factor analysis to determine the validity of a new scale variable, then reliability analysis to assess the scale's reliability. Consider different methods to create the scale, including how to handle missing data.

## Introduction

As exemplified in this summary flowchart of survey analysis



one of the first potential tasks in survey analysis is to create new variables. We do so for two reasons. First, the measurement of some concept may require that two or more variables be combined. For example, if a survey asks separate questions about the number of visits to the dentist, eye doctor, and other health care



providers, to obtain a measure of the total number of visits to all providers we need to sum these individual responses. Second, we create new variables to increase the reliability and validity of our measures over that for single questions. In this chapter we construct new variables of both types.

Another common procedure is to modify existing variables by doing recoding (binning) or using mathematical transformations (e.g., square root or log). We do not provide an illustration of these procedures here (covered in other training courses- the trainer will guide you).

Before using SPSS to accomplish these tasks, we require a firm foundation in the concepts of reliability and validity.

We begin with what might seem rather complex topics, rather than simple reports and graphics to display survey responses. This is because it is best to have all new variables created and to have studied the bivariate relationships of interest before creating any final reports and graphics to be used in a survey report. If the report is written too early, it may have to be redone because of subsequent findings.

## **Measurement And Error**

When we write questions and elicit responses, we are interested in obtaining the most accurate and precise answers possible. We never expect perfect accuracy, however, even for demographic questions, and the problems of accuracy become more severe for questions about attitudes or past behavior. It is helpful to have a framework to consider error in more depth, so we turn now to the issues underlying measurement error. A simple measurement equation to use in our examination is listed next, taken from classical test theory.

$$X = t + e$$

Here,  $X$  is the observed response to a question,  $t$  is the true value of the variable in question, and  $e$  is a random error component. In other words, every observed response is composed of two parts: a true score and error.

The true score is unobserved and is what we are trying to measure with the question and its responses. The random errors occur for a variety of reasons (fatigue, poor question wording, social desirability effects, and so on), but their existence causes the observed responses to deviate from the true score.

We will use this simple equation in our discussion of reliability and validity.

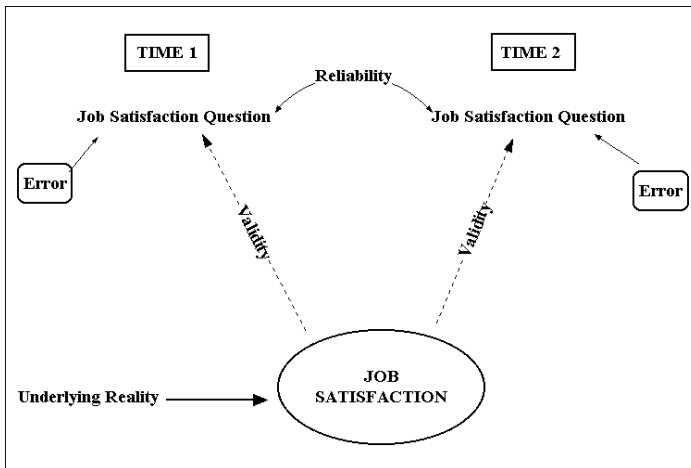
## **Reliability**

*Reliability* in survey research has the same basic meaning as it does in everyday life. It concerns consistency, or the extent to which a measuring procedure yields identical results on repeated trials. If we ask employees about their job satisfaction at time 1, and then ask them again at time 2 using the same question, the consistency of the two sets of answers is a direct measure of the reliability of the question and its response categories.

While there are several methods to measure consistency, two basic methods of estimating reliability are commonly used in surveys: the test-retest method and the internal consistency method. The retest method is the simplest and is the standard technique to measure the reliability of a single question. It requires that the same measure be collected at two points in time. In other words, we ask the same question of the same group of respondents at two times (typically in separate surveys). This is illustrated in Figure 1. The correlation between these two measurements can be interpreted as a reliability coefficient.

If there is no error, then each question will measure the true score, as can be seen from the equation above. In that case, the correlation would be a maximum (1.0), and the reliability would be perfect. Of course, we know this is never true, and the existence of random error, also depicted in Figure 1, will reduce the reliability of the question since a response will not necessarily be equivalent to its true score for a particular respondent, even if the question is ideal.

**Figure 1 The Relationship Between Reliability, Validity And Question Responses**



While intuitively appealing, the retest method is difficult and expensive to use in a survey setting, as it requires a second survey, which very few investigators in an applied setting ever complete. Furthermore, a low reliability coefficient might indicate several conditions other than an unreliable measure. Thus, for a study of employee attitudes, if there is a sufficient time lag between the two surveys, individual attitudes may change. If so, a lack of correlation indicates fundamental variation, not unreliability. Of course, the greater the time interval between surveys, the greater the likelihood of a problem of this type. Memory can also be a problem in the test-retest approach. The respondent's memory of previous answers could serve to artificially inflate the reliability coefficient.

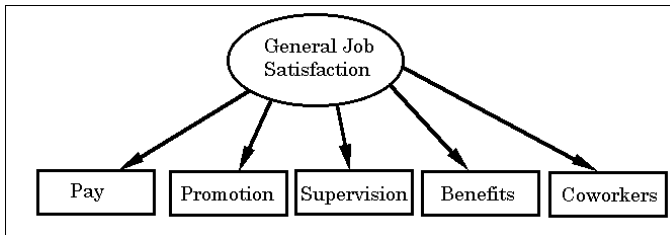
Although it can't be used to assess the reliability of a single question, the internal consistency approach is more workable for most survey researchers. It focuses on measuring several indicators of a phenomenon and evaluating their joint consistency or homogeneity. For example, we might ask three questions about various aspects of job satisfaction, and we reason that if someone is dissatisfied with his or her job, she will score low on all three items. Under this method, all the questions are asked at the same time, so it can be used in any survey. However, note that this approach requires the use of two (and preferably more) questions to measure reliability, which may be undesirable in many circumstances.

A stylized depiction of internal consistency reliability is shown in Figure 2. Here, we are using five questions about the respondent's satisfaction with various aspects of her job to measure overall job satisfaction. However, theoretically, the relationship runs in reverse: we assume there is an underlying *latent variable*, which we cannot measure (here called general job satisfaction). It directly influences the responses to the five questions on specific aspects of satisfaction. The stronger the effect of the latent variable on the questions, the higher is the internal consistency reliability.

Relating this to Figure 1, the latent variable is the underlying reality we wish to measure. The random errors of measurement mean that we can never measure it exactly with our individual questions. And, just as with the test-retest method, we need more than one measure of a concept to determine the reliability of a question. For the internal consistency method, we determine the reliability of a composite measure by using two or more questions.

A further benefit of using more than one question to measure some concept is that, almost invariably, the questions as a whole will be more reliable than any one question.

**Figure 2 Internal Consistency Reliability**



## Validity

A question must be more than reliable if it is to accurately measure some underlying concept. Thus, reliability by itself tells us nothing about whether we are truly measuring an underlying reality, even if the reliability is 1.0. A scale that always reads ten pounds too low but otherwise is error-free is thus perfectly reliable: every time I weigh myself on it, it reads ten pounds under my true weight. But each measurement is in error.

A question that measures what we intend it to is a *valid* question. Validity is a deep subject and obviously is the central concern of any measurement process.

In terms of the measurement equation, validity relates to whether the true score component is the true score of the concept we wish to measure, or something else. For example, we may use a question asking about the frequency of visits to the doctors in the past two years. If this question is not defined adequately so that some people include practitioners of alternative medicine or other health professionals in their answer, but other people exclude these categories, the question will have less than perfect validity because it will not measure precisely what we targeted.

With regard to the error term in the equation, validity is affected by non-random, or systematic error, which leads to bias. The scale that always weighs ten pounds too low exemplifies this; the scale does measure weight (instead of height or eye color), but the problem is the systematic error in that measurement. In survey research, though, it is more probable that validity will be reduced by not measuring exactly what we intend to.

Validity in Figure 1 is exemplified by the dotted lines connecting the job satisfaction latent variable with the questions. The dotted lines indicate that we are attempting to measure something called job satisfaction but are never completely certain of our success in doing so. Note how the error is not related to the underlying reality but instead directly influences the responses to the questions. Consequently, a question with no random errors would have a reliability of 1.0, but this would tell us nothing about its validity (reliability is a necessary, but not sufficient condition, for validity).

---

There is no straightforward way to measure validity as there is for reliability, though like reliability, validity is a matter of degree. Validity is not absolute since not every question can be used in every situation, as it depends on the context, population, and other factors in which a question is used. Since validity is not absolute, we do not assess the validity of an indicator but instead the validity of the *use* to which it is being put. A measure of job satisfaction for American workers may not be valid for workers in Asia, for example.

Since validity is troublesome to measure, several different types of validity have been developed. We review them briefly here.

**Face Validity.** This is the extent to which a question seems to measure what it claims to measure, just based on a close reading and study of the question. Every question should be assessed for face validity, and we normally do this without using this label to describe our review. There is no way to quantify this type of validity.

**Content Validity.** This is the extent to which a question, or set of questions, reflects a specific domain of content, body of knowledge, or specific set of tasks. It is used extensively in test construction by psychologists and educators, but less so by survey researchers, and is best used for a group of questions rather than one item. Thus, a scale composed of a group of items has content validity if it adequately represents the universe of potential questions that could be used to measure a specific concept. In our example in Figure 2, these five questions would have content validity if they were judged to be a representative sample that covers the potential components of job satisfaction. If, instead, an area were not represented (say, how interesting or challenging the work is), then the overall measure would have lower content validity. There is no direct method to quantify content validity, but it should always be assessed when creating multi-item scales by thinking carefully about the concept to be measured, or consulting with experts on the topic.

**Criterion Validity.** This type of validity is the most practical and the most straightforward to measure, at least in some surveys. Criterion validity concerns how well a question or scale is associated with an external criterion. Thus, an external measure is used to validate a question on a survey. If a written exam in a physical education class accurately predicts one's ability to play volleyball, then the exam has criterion validity. The magnitude of the correlation between the two things is a measure of the level of validity. Or, if we believe that a measure of overall customer satisfaction is valid, then it should predict future buying behavior. So if we find a high, positive correlation between the two, we would consider the overall satisfaction question to be reasonably valid for its intended job (predicting future purchases).

Criterion validity is rather atheoretical in that it is determined by the empirical association between a question and a criterion. However, in practice, the criterion is always chosen so that it makes both theoretical and practical sense to use it for validation. Thus, to validate a question about income, we could use tax returns. To validate a question about voting, we would use election records. We attempt to find a criterion that is measured at least as accurately as the question being studied; otherwise, a lower association may be caused by measurement error in the criterion.

There are two types of criterion validity. *Concurrent validity* is assessed by measuring the criterion and the behavior or attitude at the same point in time. *Predictive validity* is assessed by measuring the criterion at a later time, as for future buying of a product.

**Construct Validity.** This type is the most sophisticated form of validity, and it assesses the extent to which a particular question or questions relate to other questions, given a theoretically specified set of hypotheses relating the underlying constructs being measured. The hypotheses specify the relationship we should find between the questions. Criterion validity is concerned with how well a measure performed; now we are concerned with *why* a measure works well. Here's an example. Let's say we expect that people who are more conservative will predictably favor certain types of laissez-faire economic policies and a strong



military. We develop a question to measure conservatism, and then correlate it with measures of economic policy and military support. If those respondents who score high on our conservatism measure are, indeed, more likely to favor a strong military and prefer less government intervention in the economy, then our conservatism measure has construct validity.

A variant of construct validity uses factor analysis to examine a group of items that constitute a scale to determine whether they form one construct. This is the approach we take in this chapter.

You should consider the face and content validity for any question, even if only crudely. It is difficult in most practical applications to estimate construct validity, but it may be possible to assess the criterion validity of several questions on many surveys, given an appropriate criterion.

To conclude our review of reliability and validity, try to identify potential problems of reliability and validity with these questions.

1. What is your age? \_\_\_\_\_

2. What is your income?

- Under €20,000
- €20,000 to €29,999
- €30,000 to €39,999
- and so forth ...

3. What is your overall job satisfaction?

(Measured on a seven point scale from “Extremely satisfied” to “Extremely dissatisfied.”)

4. Do you plan to buy another laptop for your home office?

- Yes
- No

## Indices and Scales

The nomenclature in survey research can be confusing. For example, the word “scale” is used to refer to two very different concepts. A Likert scale question uses an ordered set of response alternatives ranging from strongly disagree to strongly agree. On the other hand, a group of items combined to create a composite measure of some concept is also referred to as a scale.

To be precise, it would be best to label the responses for a single question a *response scale*, where the word “response” indicates that the scale refers to an individual question. The word “scale” by itself should be reserved for items constructed from two or more questions.

Then, to further complicate matters, there are two types of variables that can be constructed from two or more existing questions. Both are commonly used in survey research. One is called an *index*, the other, a scale, and the difference between them is theoretical, not in how they are constructed.

- 1) An index is a composite measure that sums together a set of question responses to determine the level of some theoretical construct. It is not caused by a latent variable, as depicted for general job

---

satisfaction in Figure 2. Instead, the individual items directly influence the composite measure (the arrow of causation goes in the opposite direction). Sometimes the creation of an index is clear-cut. For example, if we have two questions asking about the number of PCs and iPads purchased, we can sum them together to get an index of the total number of computers bought. However, other times, there is a less direct relationship between the individual questions and the composite measure. Social scientists have developed the concept of socioeconomic status (SES) to measure a person's overall status in society. It is an index composed of education, income, and a measure of the status of one's occupation. These measures are combined in a sophisticated method to construct SES. But, again, these three measures determine SES, rather than SES somehow determining one's education, etc.

- 2) A scale uses two or more questions to form a composite measure of a unitary theoretical concept (see Spector, 1992 for a comprehensive introduction). Underlying this is the theoretical relationship depicted in Figure 2. Here, the individual items *are* determined or caused by the latent underlying concept: they are imperfect measures of it. Scales have been developed to measure just about everything, it would seem, including job satisfaction, self-esteem, organizational commitment, and so forth. Similarly, verbal and maths scores are, in a real sense, two scales composed of hundreds of items to measure those two types of abilities. (The argument over such tests is not about only about reliability but also about validity: does it measure what the test developers and users claim it does.) Scales are developed because it is hoped that they will do a better job of measurement than any one question. In other words, a scale will be a more reliable and valid measure of a concept than a single question. It can then be used in place of the individual items in further analyses. PSYCHOMETRICS is a good example.

Many questionnaires lend themselves naturally to scale development because they contain several questions about one underlying concept. (Scales are typically developed to measure attitudes, but there is no theoretical reason why they cannot measure behavior.) Many surveys ask several questions about a single concept. For example, a survey of hospital patients might ask several questions about satisfaction with medical care. The observed responses to the items can be conceptualized as being caused by an underlying, unmeasured construct of satisfaction with the medical care.

As we saw in Figure 2, the five questions are all concerned with job satisfaction. If the responses to these questions are all related to or caused by some underlying, unmeasured concept of general job satisfaction, then we should be able to create a valid scale from them. The five items should also not be influenced by any other constructs (such as overall satisfaction with one's life or spouse).

It turns out that measuring the reliability, and to a lesser extent, the validity, of a group of items forming a scale is more readily accomplished than measuring these properties for the individual items.

Measuring the validity of an index is unnecessary because, by definition, the set of items defines or determines the composite variable. Still, assessing the validity of the individual questions used in the index can be a worthwhile task. And, it is possible to disagree over whether a particular group of questions is the best set to create an index. The reliability of an index can be measured since random measurement error will still affect it.

We turn next to the construction of an index.

## ***Constructing an Index***

We will create an index that measures how much a respondent uses the Internet for political purposes. To do so, we utilize five individual questions that ask about specific types of Internet usage. This will be an example of an index that is unproblematic to create because it is clear that the individual items relate directly to the theoretical concept we are trying to measure.



To accomplish the task, we use an old survey conducted by the Survey Research Center of the University of Maryland in 1996. Although it is probably not critical for this example, note that the data were collected via a telephone interview. Mode of data collection should always be kept in mind when doing analyses. The five questions, and the variable names used in the data file, are listed next.

1. People use the Internet for many purposes. One purpose can be to learn about government and politics. During the last year, have you used the Internet to find out what the government or a particular official is doing? (*intgovt*)
2. During the last year, have you used the Internet to contact a public official or candidate for office? (*intcand*)
3. During the last year, have you used the Internet to express your views about politics or government to others? (e.g. by sending e-mail or posting a message to a newsgroup) (*intpol*)
4. During the last year, have you used it (the Internet) to learn about political issues (such as gun control, taxes, etc)? (*intlearn*)
5. (During the last year,) have you used it (the Internet) just to browse for political information with no specific use in mind? (*intbrows*)

The response scale for each question was a yes/no dichotomy (yes=1, no=2). It seems reasonable to add these items together to create a summary measure of how much a respondent uses the Internet for political purposes. The index will not be a measure of frequency of use since the questions don't ask about that.

We now open SPSS and the data file.

### **A Note Concerning Data Files, Variable Names and Labels in SPSS**

In this course, we want the variables to appear in dialog boxes in their file (questionnaire) order, not alphabetical order. This will make them easier to match up with the questionnaire. Second, for the same reason, we want the variable names, not labels, to be displayed. We also want to display numbers in standard format, not scientific notation.

- Click the **Start** button, then **Programs**, then **SPSS for Windows**
- Click **Edit...Options**, then click the **General** tab (not shown)
- Click the **Display names** option button in the Variable Lists section
- Click the **File** option button in the Variable Lists section
- Click the **No scientific notation for small numbers in tables** checkbox

In addition, when working with survey data, it is best to see both the labels *and* the actual values in the data for responses in tables. This makes it easier to understand how to sum or recode variables.

- Click on the **Output Labels** tab
- In the dropdown list for Variable values in labels, change it to **Values and Labels**
- Click **OK**

Course files are assumed to be located in the c:\Train\ directory. (containing this guide). If you are running SPSS Server (you can check by clicking File...Switch Server from within SPSS), then files used with SPSS should be copied to a directory that can be accessed (mapped) from the server.

To access the data from SPSS:

- Click on **File...Open...Data**
- Move to the c:\Train\SurveyAnalysis directory
- Double-click on the **MARYLAND\_1271.SAV** file to open it



The data file contains 1013 cases. However, the data must be weighted to match population distributions for analysis, and the resulting weighted file has 1000 cases. (Although ideally the weighted data file should be the same size as the unweighted file, small differences like this occur in practice.)

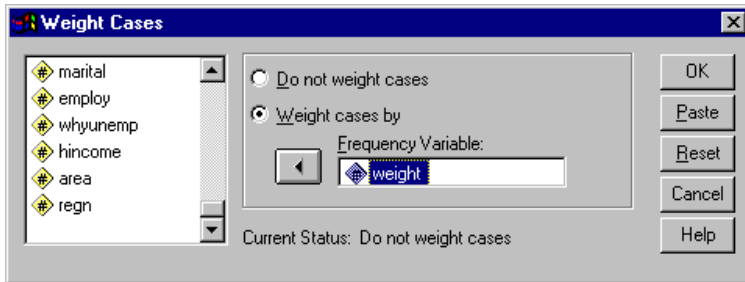
To weight the data:

Click **Data...Weight Cases**

In the Weight Cases dialog box, click the **Weight cases by** option button

Move **weight** (the last variable listed) into the **Frequency Variable** box

**Figure 3 Weight Cases Dialog Box**



Click on **OK**

When the file is weighted, the words “Weight On” appear in the right corner of the Status Bar of the Data Editor window.

Now we can request frequency distributions for each variable.

Click **Analyze...Descriptive Statistics...Frequencies**

Move **intgovt**, **intcand**, **intpol**, **intlearn**, and **intbrows** into the **Variables** list box (not shown)

Click on **OK**

First, from the summary table (not shown) we see that there is a great deal of missing data (795 cases). This is because not every respondent was asked these questions, so this type of missing data is not a concern (although it does greatly reduce statistical power).

Figure 4 displays two of the five frequency tables. The percentage of respondents engaging in various political related activities on the Internet varies from 31.6% (browsing for political information) to a low of 10% (contacting a public official). We use the valid percent column in the Frequencies table for these statistics.

**Figure 4 Frequency Distributions for Intgovt and Intcand**

USED INTERNET FIND WHAT GVNT DO LAST YR					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 YES	54	5.4	26.4	26.4
	2 NO	151	15.1	73.6	100.0
	Total	205	20.5	100.0	
Missing	8 DK	1	.1		
	System	794	79.5		
	Total	795	79.5		
Total		1000	100.0		

USED INTERNET TO CONTACT PUBLIC OFFICIAL					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 YES	21	2.1	10.0	10.0
	2 NO	185	18.5	90.0	100.0
	Total	205	20.5	100.0	
Missing	System	794	79.5		
Total		1000	100.0		

An index created from these five variables may not represent all the ways in which the Internet can be used for political purposes, but it certainly would appear to be a reasonable measure of this concept, since five different types of uses will be included in the index. When creating an index, you need to think carefully about exactly what it will measure (in truth, you should be thinking about this when writing the questionnaire).

Normally, an index is created by summing together the individual variables. We can't immediately do so, though, for two reasons. The current coding of the variables, with a "no" response coded as a 2, would create a summed score where lower scores indicate more Internet use. This would be awkward to use in reports and tables, so we need to recode the variables so that a "no" response is given a value of 0.

The second reason concerns missing data, caused by a don't know response or a refusal to answer the question. In this set of five questions, there is only one don't know response (the "DK" for intgovt). (Normally these values are defined as user-missing, but if they are not, they will be included in the summed scores.)

When variables are summed, a missing value on one or more of them will cause the new index to be coded as system-missing by default in SPSS. This may be appropriate in many instances, especially when several of the variables have a missing response. But if an index is composed of several items, but only one variable has missing data for many respondents, then many cases will be lost when creating the index. Two common methods to retain these cases are:

- 1) Replace the missing data with the mean, or median, value for that variable.
- 2) Sum the variables with valid values for each respondent, and then multiply by a correction factor to take into account the one missing variable (if there are five items, and one is missing, multiply by 5/4).

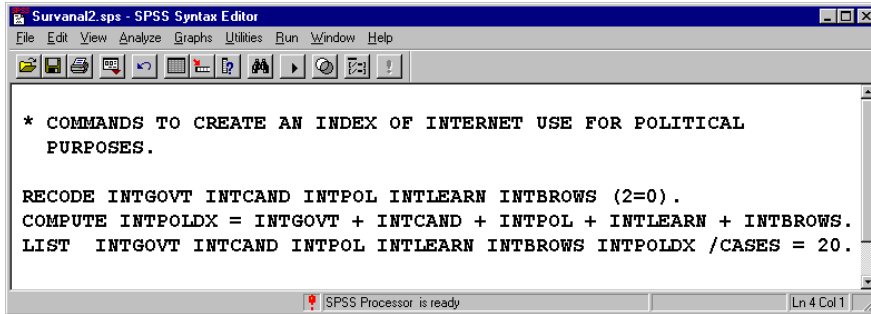
Again, these methods are normally used when only one or two of the variables have missing values. There is no point in using them when creating the Internet index because there is so little missing data.

To reduce the time involved in creating the index, we have supplied a syntax file with the appropriate commands.

Click **File...Open...Syntax**

In the directory **c:\Train\**, double-click on the file **SURVANAL2.SPS**

**Figure 5 Syntax to Create Index Variable**



```
Survanal2.sps - SPSS Syntax Editor
File Edit View Analyze Graphs Utilities Run Window Help
* COMMANDS TO CREATE AN INDEX OF INTERNET USE FOR POLITICAL
PURPOSES .
RECODE INTGOVT INTCAND INTPOL INTLEARN INTBROWS (2=0) .
COMPUTE INTPOLDX = INTGOVT + INTCAND + INTPOL + INTLEARN + INTBROWS .
LIST INTGOVT INTCAND INTPOL INTLEARN INTBROWS INTPOLDX /CASES = 20 .
Ln 4 Col 1
```

The commands in this file change the value of 2 to 0 for the five variables, compute the new index, and then use the List command to view the variables used to create the index and the index itself for the first 20 cases in the file. This last step is absolutely necessary to insure that the index was constructed correctly. All of this can be done from the dialog boxes, but it saves time to have the commands already created.

Note that we wrote a mathematical expression to add the variables together rather than make use of the Sum function. This is because that function by default will sum together whatever valid values exist for a list of variables, so a respondent with most data missing will still obtain a valid sum for the index.

Now we run the commands.

Highlight all **three commands**

Click on the **Run button**  or **Run...Selection** from the menus

Figure 6 Output from List Showing Index Variable Intpoldx

INTGOVT	INTCAND	INTPOL	INTLEARN	INTBROWS	INTPOLDX
1	0	0	1	1	3.00
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
0	0	1	0	0	1.00
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
0	0	0	0	0	.00

Number of cases read: 20    Number of cases listed: 20

We observe that the index variable `intpoldx` has been successfully created. The first respondent answered yes to the questions `intgovt`, `intlearn`, and `intbrows`, and so has an index score of 3.

As with any variable, we would next examine its distribution and univariate statistics. We will briefly do this for `intpoldx`.

Click **Analyze...Descriptive Statistics...Frequencies**

Click the **Reset** button

Place **intpoldx** in the **Variables** box

Click the **Statistics** button, then on the **Mean, Median, and Std. deviation** checkboxes (not shown)

Click on **Continue**, then on **OK**

We see that there are 204 valid cases, as expected. The mean is rather low, 1.11, which means that the Internet is not used much for political purposes. Over 50% of the respondents have an index score of 0, which indicates that they don't use the Internet for any of these activities. Conversely, only seven people answered yes to all five questions.

**Figure 7 Frequency Distribution and Univariate Statistics for Intpoldx**

**Statistics**

INTPOLDX		
N	Valid	204
	Missing	796
Mean		1.1125
Median		.0000
Std. Deviation		1.4228

**INTPOLDX**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	109	10.9	53.2	53.2
	1.00	23	2.3	11.2	64.4
	2.00	35	3.5	17.4	81.7
	3.00	22	2.2	11.0	92.7
	4.00	8	.8	4.0	96.8
	5.00	7	.7	3.2	100.0
	Total		204	20.4	100.0
Missing	System	796	79.6		
	Total	1000	100.0		

With the index created, it can now be used for reporting and analysis.

## Constructing a Scale

We next construct a scale variable using a customer satisfaction survey completed for SPSS. Specifically, we use several questions that asked about various aspects of the performance and quality of work of the respondent’s sales representative at SPSS. The items about the sales rep are Questions 18 through 24. They are answered on a scale of 1 to 5, where 1=Strongly Agree and 5=Strongly Disagree, so this is a classic Likert (response) scale. The SPSS data file (SPSS\_CUST.SAV) contains 955 cases and the data were collected via a mail survey.

With several questions about the same topic (sales rep performance), we hypothesize that there is one latent variable that is determining the observed responses to these seven questions. If this is true, we can create a scale to increase the reliability and validity of our measure of sales rep performance.

To do this we use factor analysis, which is a method to determine whether clusters of variables exist in a file. However, we must be careful in our understanding of what factor analysis can and cannot do. Factor analysis by itself cannot answer the question of whether these seven items are truly a valid measure of job performance. It can only tell us, as per Figure 2, whether there is an underlying factor—whatever it is—that influences the responses. To put it another way, factor analysis finds clusters of interrelated variables that, if strongly associated, are probably all measuring various aspects of the same concept, *whatever that concept is*. You, the survey researcher, must determine what that concept is.

Admittedly, in many instances, that decision is not difficult, as is very likely true that Questions 18 through 24 are truly measuring some aspect of the sales rep’s job performance. Determining this is a matter of content validity. And, if factor analysis shows that the variables are valid measures of an underlying concept, that is a type of construct validity.

## Principles of Factor Analysis

Factor analysis operates on the correlation matrix of the variables to be factored. The basic rationale is that the variables are correlated because they share one or more common components, and if they didn't correlate there would be no need of or use in performing factor analysis.

Employing correlations implies that we are using variables measured on an interval/ratio scale, one of the assumptions of factor analysis. However, the use of factor analysis with five, six, or seven point response scales is very common with survey research data, and experience has shown that the true ordinal nature of these response scales has little detrimental effect on the factor results (one should be very cautious using factor analysis for response scales with fewer than five points). It is possible to use nonlinear principal components to investigate the data structure for ordinal variables, but this topic goes beyond the scope of this course.

Since factor analysis is a multivariate statistical method, the rule of thumb for sample size is that there should be at least 10 to 25 times as many observations as there are variables used in the analysis. This is because factor is based on correlations, and for  $p$  variables there are  $p*(p-1)/2$  pairwise correlations. Think of this as a desirable goal and not a formal requirement. Given the sample size of most surveys, it should be no problem meeting this requirement.

Let's open the SPSS data file and check the correlations among the variables.

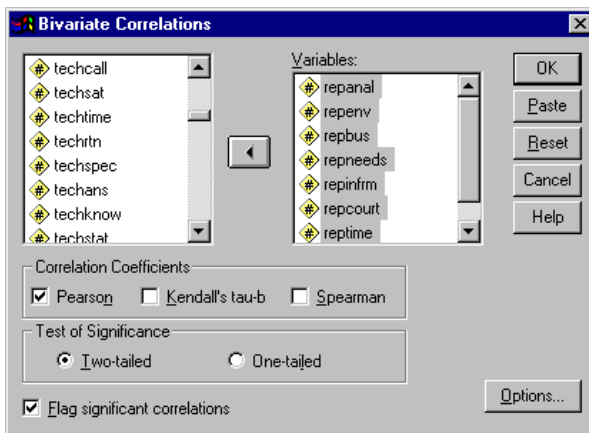
Click **File...Open...Data**

Click on **SPSS\_CUST.SAV** to select it, then click the **Open** button (don't save the current data file when asked)

After the file is open, click on **Analyze...Correlate...Bivariate**

Place the variables **repanal** to **reptime** in the **Variables** list box

**Figure 8 Bivariate Correlations Dialog Box**



Click **OK** to create the correlations

The resulting table of correlations has been edited to display only the Pearson correlation coefficient and the number of valid cases for that statistic. Looking at either the upper or lower triangle (since the table is symmetric), we note that all the correlations are positive and fairly large, ranging from .290 to .719, with most in the .40 to .50 range. This is reassuring and indicates that factor analysis is a reasonable technique to try.

**Figure 9 Correlations Among The Sales Rep Items**

Cells contain correlation and valid number of cases.		Correlations						
		Sales rep undrstrnds my stat data anal needs	Sales rep undrstrnds computng envrmt	Sales rep undrstrnds your business/organ	Sales rep relates prods to my needs	Sales rep informs about all prods & svcs	Sales rep treats customer w/courtesy	Sales rep gives info in right amt of time
Sales rep undrstrnds my stat data anal needs	Pearson Correlation N	1	.623**	.689**	.693**	.478**	.499**	.480**
		585	555	510	540	575	580	580
Sales rep undrstrnds computng envrmt	Pearson Correlation N	.623**	1	.702**	.593**	.534**	.418**	.523**
		555	575	520	535	555	570	570
Sales rep undrstrnds your business/organ	Pearson Correlation N	.689**	.702**	1	.719**	.507**	.290**	.398**
		510	520	530	515	520	525	525
Sales rep relates prods to my needs	Pearson Correlation N	.693**	.593**	.719**	1	.561**	.353**	.390**
		540	535	515	555	550	545	545
Sales rep informs about all prods & svcs	Pearson Correlation N	.478**	.534**	.507**	.561**	1	.576**	.544**
		575	555	520	550	610	600	600
Sales rep treats customer w/courtesy	Pearson Correlation N	.499**	.418**	.290**	.353**	.576**	1	.605**
		580	570	525	545	600	625	620
Sales rep gives info in right amt of time	Pearson Correlation N	.480**	.523**	.398**	.390**	.544**	.605**	1
		580	570	525	545	600	620	625

\*\* Correlation is significant at the 0.01 level (2-tailed).

There are two steps in a factor analysis.

**First, factors must be extracted from the variables.** With most survey data, researchers use either principal components extraction or principal axis factoring. The former attempts to account for the maximum amount of variation in a set of variables. The latter attempts to account for correlations between the variables. There is much overlap between these two extraction methods and they usually yield similar results. However, since we are interested in understanding the relationships between the variables (do they form one factor?), we recommend using principal axis factoring. This is especially true if the items were not written expressly to be part of a scale.

**On the other hand, much customer satisfaction data can have high correlations between the various questions (say .80 or above), which means there is likely to be multicollinearity in the data.** In that case, most factor methods can't be used, but principal components can.

The number of factors extracted is determined, by default, by all those factors that explain more variance than a single item (technically, this implies an eigenvalue greater than 1.0). Other criteria are possible as you become more skilled in the use of factor analysis.

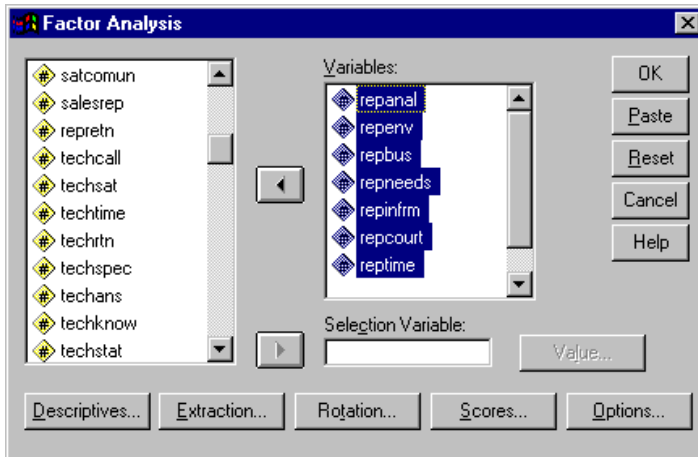
**Second, after determination of the number of factors, the factors are rotated,** though often kept orthogonal, to maximize the size of the coefficients—called *loadings*—that relate the factors to the variables. This is done for ease of interpretation, as the ideal situation is one in which a variable loads highly on one factor and very low on all other factors. Principal components solutions don't have to be rotated, for technical reasons beyond the scope of this discussion, although they often are. The most popular rotation choice is varimax, which attempts to simplify interpretation by maximizing the variances of the variable loadings on each factor (i.e., it tries to simplify the factors). When studying scales it is strongly recommended that this rotation be used.

## Running Factor Analysis

We are now ready to run a factor analysis on these seven questions. Other technical issues about factor analysis will be mentioned as we run the procedure.

Click on **Analyze...Data Reduction...Factor**  
 Move **repanal** to **reptime** into the **Variables** list box

Figure 10 Factor Analysis Main Dialog Box



We need to specify at least the Extraction and Rotation methods to complete the analysis.

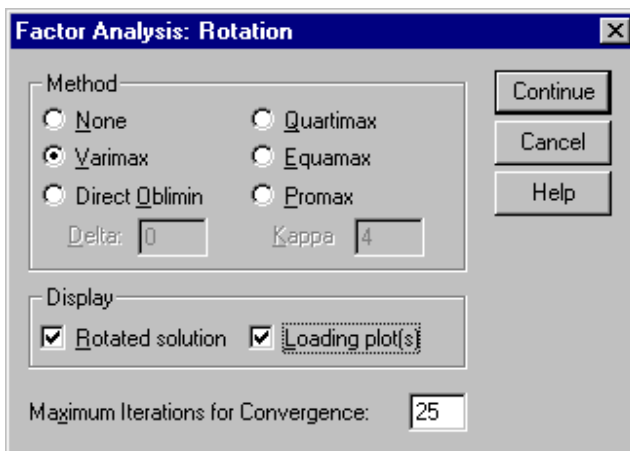
Click the **Extraction** pushbutton

In the drop-down list for **Method**, choose **Principal axis factoring**, then click **Continue** (not shown)

Click the **Rotation** pushbutton, then click the **Varimax** option button

Click the **Loading plots** checkbox; this will display a plot of the factor loadings

Figure 11 Rotation Subdialog Box



Click **Continue**, then click the **Options** pushbutton

By default, Factor Analysis uses listwise deletion to handle missing data. This means that if a case has missing data for one or more variables, it will be removed from the analysis. Looking again at the correlations in Figure 9, we see that there is a substantial amount of missing data for these seven questions. However, only those respondents who spoke to an SPSS sales rep in the last year answered these questions, and by running Frequencies on Question 16 we can learn that this includes 645 respondents. Nevertheless, the number of valid responses to the individual items varies from 625 to 530, with the greatest amount of missing data for the question about understanding business needs.

The safest approach is to use the default listwise deletion, but this will reduce the number of cases in the analysis to 530 or less (the actual value will be 490 and will be reported if you request Univariate



descriptives). This is a substantial drop from 645 (about 24%). Even if you find this acceptable, it implies that any scale we create from these questions will have a valid value for only 490 of the respondents, and this could seriously limit any further analysis you do with the scale. You can investigate whether the respondents dropped from the analysis are different on important variables from the ones remaining in the analysis.

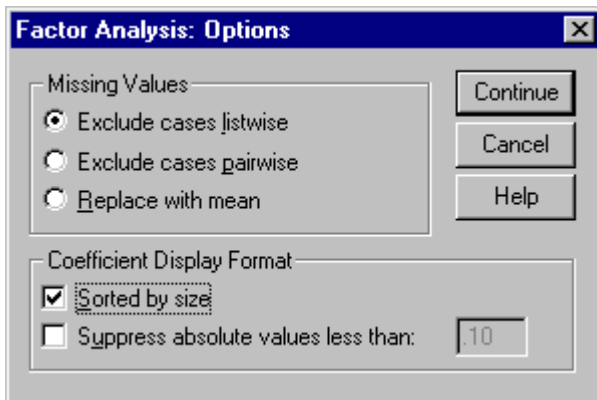
Using the two other options isn't recommended except when the amount of missing data is small (but in that case, listwise deletion won't reduce the file size significantly, so you could use that default option anyway). When data are missing at random (see the chapter on missing data), you could substitute the mean, although that has the somewhat undesirable property of reducing the variance of the items, which can modify the factor structure.

We return to this topic when creating the scale.

Click the **Sorted by size** checkbox in the **Coefficient Display Format** area

The Sorted by Size option will have SPSS list the variables in descending order by their loading coefficients on the factor for which they load highest. This makes it very easy to see which variables relate to which factors.

**Figure 12 Options Subdialog Box**



Click on **Continue**, then **OK** to run the analysis

In the output from the procedure, we ignore the Communalities table and turn to the Total Variance Explained table. The variables are not represented in this table, only the factors. The Initial Eigenvalues section contains all seven eigenvalues for the seven possible factors (there can be as many factors as variables). Only two of these factors have eigenvalues above 1 (the second just barely). These two factors will be extracted in the first part of the analysis. The third factor's eigenvalue was not close to 1, so it appears that a two-factor solution is reasonable.

**Figure 13 Total Variance Explained Table**

**Total Variance Explained**

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.334	61.914	61.914	4.009	57.270	57.270	2.559	36.554	36.554
2	1.017	14.532	76.446	.693	9.896	67.166	2.143	30.612	67.166
3	.477	6.811	83.257						
4	.408	5.829	89.087						
5	.335	4.784	93.871						
6	.232	3.311	97.182						
7	.197	2.818	100.000						

Extraction Method: Principal Axis Factoring.

Of course, we’ve already run into a potential dilemma. We hypothesized that the seven questions form one common scale, and the factor analysis indicates that this might not be true, as two factors were extracted.

In the Extraction Sums of Squared Loadings section—which uses only the extracted factors—we see that the first factor explains over 57% of the variance, which is quite large. The second doesn’t explain that much, less than 10%. Together they account for 67% of the variance, which is quite good.

After the two factors are rotated, the Rotation Sums of Squared Loadings shows that they account for the same amount of variance together, but the amount of explained variance for each has been almost equalized. Factor analysis techniques have no unique orientation of their axes, so the rotated solution is as viable as the unrotated solution. The equality of variances for the factors is further evidence that we are not dealing with a unidimensional scale measuring sales rep performance.

Scroll by the Factor Matrix table and examine the Rotated Factor Matrix output. The numbers in the table are the loadings of each variable on the two factors, after a varimax rotation has been applied. The loading is the correlation between a variable and the (unobserved) factor. In the context of Figure 2, loadings are a measure of the strength of the effect of the theoretical concept on the observed indicators.

**Figure 14 Rotated Factor Matrix**

**Rotated Factor Matrix<sup>a</sup>**

	Factor	
	1	2
Sales rep understands your business/organ	.851	.222
Sales rep understands my stat data anal needs	.767	.362
Sales rep relates prods to my needs	.755	.288
Sales rep understands computng envmt	.621	.471
Sales rep treats customer w/courtesy	.225	.760
Sales rep gives info in right amt of time	.270	.746
Sales rep informs about all prods & svcs	.408	.722

Extraction Method: Principal Axis Factoring.  
Rotation Method: Varimax with Kaiser Normalization

a. Rotation converged in 3 iterations.

---

### Technical Note

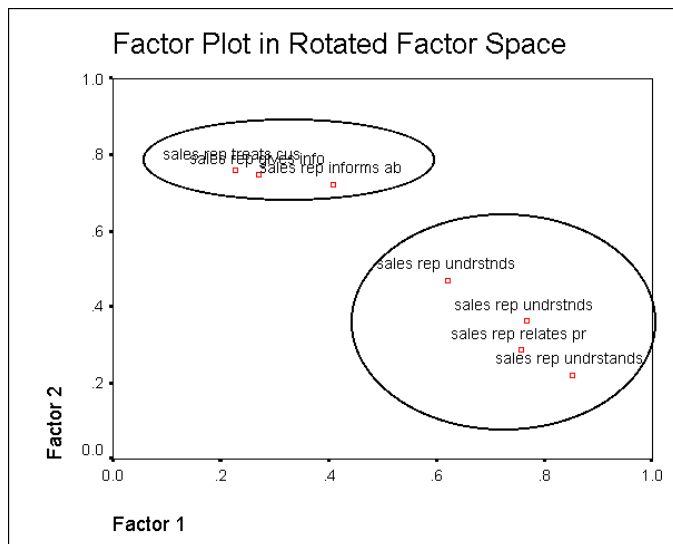
It is important to check the footnotes for the Factor Matrix table and the Rotated Factor Matrix table that discuss whether, respectively, the extraction was successful and the rotation converged. If more than the default number of iterations is required for either or both, additional iterations should be requested until a solution is obtained.

The questions most associated with factor 1 are the first four listed in descending order by loading. They are Questions 18 through 21. All have high loadings above .50 (which is roughly the lower limit that we use for determining whether a variable loads on a factor), with “Understands your business/organization” having the highest loading. All these four have lower loadings on the second factor, although Question 19 about sales rep understanding the computing environment has a loading of .471 on factor 2.

The second factor (again, we expected not to see a second factor if the scale is unidimensional) has questions 22 through 24 loading highly. All the loadings are above .70 and very similar. When creating a scale, one looks to see whether the individual items measure the concept equally, which is the case for factor 2.

The factor structure is graphically depicted in Figure 15, the Factor Plot. The graph has been modified to zoom in on the upper right quadrant where the variables are plotted according to their loading on the two factors. The ovals have been added to emphasize how two distinct factors have been extracted. If there were only one dominant factor it would not be possible to separate the variables in this two dimensional space, but we can see that the questions are reasonably separated into two distinct factors with no overlap.

**Figure 15 Factor Loading Plot**



### Factor Interpretation

The next task is to interpret the factors, which is somewhat of an art. For factor 1, given that the word “understanding” is used in three of the items, and that all the questions relate explicitly or implicitly to an SPSS user’s various “needs,” we might describe the factor as measuring the sales representative’s ability to understand a customer’s needs and organization.



Two items on the second factor are concerned with giving the customer information, the third with courtesy. All ask about actions taken by the sales representative rather than an abstract concept of understanding. We might label this factor a measure of the sales representative's quality of customer interaction and service.

In this light, it becomes less surprising that factor analysis extracted two factors. One deals with the sales rep's ability to understand the customer (three of the questions on factor 1 begin with that word), the other with service. This implies that the seven questions together do not form a unidimensional scale that measures the sales rep's performance. You can now understand why we assess validity before reliability. Although factor analysis has made it very probable that the seven items do not measure one unitary theoretical concept, it has not truly assessed the validity of each factor. What each factor measures, and, more to the point, how well it does this, remains an open question. For example, there are different wordings we could have used for these seven items, and the wordings might be better or worse measures of the underlying concept. And there is no guarantee that the items ask about all aspects of the performance of a sales rep.

Still, the factor analysis has prevented us from making an error by using all seven items together. Put another way, finding that items load together on a factor is a necessary, but not sufficient, indication that the scale constructed with them will be valid.

Next we turn to assessing the reliability of the scale.#

## **Reliability of Scales**

We found two potential scales to measure the performance of the SPSS sales representatives. Now we want to calculate each scale's reliability.

Since we only have measurements at one point in time, we will assess reliability with an *internal consistency* statistic, Cronbach's alpha. Internal consistency is based on the idea that items comprising a scale should show high levels of internal consistency, which for alpha reduces to high inter-item correlations. The higher the correlations among the items, the greater the alpha. High correlations imply that high (or low) scores on one question are associated with high (or low) scores on other questions. In the context of our example, respondents who strongly agree that the sales rep understands their business are also likely to agree that the sales rep understands their computing environment, if reliability is high (because both are affected in a similar manner by the underlying latent variable).

Cronbach's alpha varies from 0 to 1. It has a number of possible interpretations. One is related to the variance of the items. The variability of the responses to a set of questions is due to actual (true score) variation, plus random errors (recall the measurement equation). Alpha divides the total variance into these two components, so that

$$\text{Alpha} = 1 - \text{error variance}$$

Thus, alpha can be viewed as the true variance in the scale.

Perhaps the most theoretically meaningful interpretation is to consider alpha the correlation between the actual scale and a hypothetical alternative scale of the same length, drawn from the universe of all possible questions about the underlying construct. This version makes plain the idea that the scale we have is only one of many possible alternative scales.

Alpha can be calculated from either the covariance or correlation matrix. Although the covariance formula is more fundamental, given the interpretation above, we present the correlational formula since most people are more familiar with correlations.

$$\alpha = \frac{(\text{Number of items})(\text{average correlation among items})}{1 + (\text{average correlation among items})(\text{Number of items} - 1)}$$

If the average correlation among the items is held constant, then alpha increases as the number of items increases. In other words, a scale with more items, all things being equal, should have a higher reliability.

Since alpha is related to scale length, rules of thumb for what constitutes an acceptable level of reliability should consider length. For short scales of five items or less, a minimum recommended level of alpha is often .70. Since a correlation between two scores squared is a measure of shared variance, an alpha of .70 means that the scale shares about half of its variance with a hypothetical alternative scale (refer to the second interpretation of alpha). As scale length increases, or as the reliability of a scale becomes critical, alpha should probably be .80 or above.

## Running a Reliability Analysis

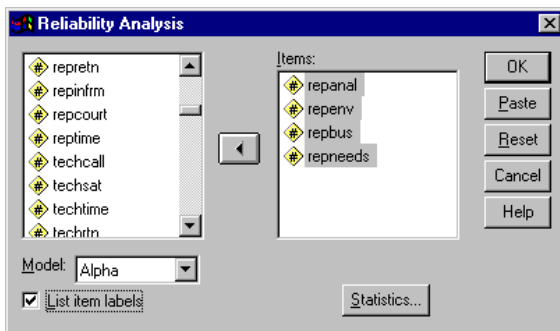
The SPSS Reliability procedure calculates alpha plus other helpful statistics. First we test the variables on the first factor.

Click **Analyze...Scale...Reliability Analysis**

Move the variables **repanal** to **reneeds** into the **Items** list box

Click on the **List item labels** checkbox

Figure 16 Reliability Analysis Dialog Box



The default model is alpha, the one we want. Only one scale can be tested at a time using the dialog boxes.

Click the **Statistics** pushbutton

Click on **Item**, **Scale**, and **Scale if item deleted** checkboxes

Click on the **Means** and **Variances** checkboxes in the **Summaries** area (not shown)

Click on **Continue**, then **OK**

The output from reliability is in text, not pivot table, format. The first section, shown in Figure 17, lists summary measures for the items and the overall scale.

The four items all have about the same mean and standard deviation. These are desirable qualities for the items comprising a scale. Actually, ideal items would have a mean near the midpoint of the response scale (5, in this case), and reasonably large standard deviations. The Max/Min ratios for the item means and variances are quite acceptable.

Notice that only 500 respondents are included in the analysis because not everyone who talked to a sales rep answered all four questions, as we noted above when doing the factor analysis.

**Figure 17 Item and Scale Summary Statistics**

RELIABILITY ANALYSIS - SCALE (ALPHA)						
1.	REPANAL	Sales rep understands my stat data anal ne				
2.	REPENV	Sales rep understands computng envmt				
3.	REPBUS	Sales rep understands your business/organ				
4.	REPNEEDS	Sales rep relates prods to my needs				
		Mean	Std Dev	Cases		
1.	REPANAL	2.3000	.8784	500.0		
2.	REPENV	2.3500	1.0816	500.0		
3.	REPBUS	2.5500	1.0342	500.0		
4.	REPNEEDS	2.6300	1.0561	500.0		
	N of Cases =	500.0				
		Mean	Variance	Std Dev	N of Variables	
Statistics for	Scale	9.8300	12.2857	3.5051	4	
Item Means		Mean	Minimum	Maximum	Range	Max/Min
		2.4575	2.3000	2.6300	.3300	1.1435
		Variance				
		.0249				
Item Variances		Mean	Minimum	Maximum	Range	Max/Min
		1.0316	.7715	1.1698	.3983	1.5162
		Variance				
		.0317				

If we construct the scale by adding all four items together, the scale will have a mean of 9.83 with a standard deviation of 3.51. Next we turn to the reliability statistics.

Moving to the bottom of the output (Figure 18), we see that alpha is quite large, .885, especially for a four-item scale. This is strong evidence that these items form a reliable measure of the sales rep’s ability to understand a customer’s business needs, which is quite encouraging.

Two additional statistics can be helpful. The “Alpha if Item Deleted” column tells us what the reliability would be of a (three-item) scale if this item were deleted from the current scale. If alpha increases when an item is deleted, that is one indication that the item may not be necessary for scale construction. However, theoretical and practical considerations should also influence the decision to drop an item. In this case, alpha gets smaller, not larger, as items are dropped.

**Figure 18 Item-total Statistics and Cronbach’s Alpha**

Item-total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
REPANAL	7.5300	7.6444	.7967	.6421	.8417
REPENV	7.4800	7.1038	.6959	.5275	.8757
REPBUS	7.2800	6.8954	.7955	.6338	.8351
REPNEEDS	7.2000	7.0541	.7337	.5988	.8597
Reliability Coefficients      4 items					
Alpha = .8855                      Standardized item alpha = .8894					

The “Corrected Item-Total Correlation” is essentially the correlation between the item in each row and the other three items in the scale. A rule of thumb is that it should be greater than .50 to retain an item, and all four items easily meet this criterion.

Overall, we have solid indications that questions 18 through 21 form a reliable scale that measures one aspect of a sales rep’s performance. In combination with the factor analysis that determined the validity of using these items as a scale, we can now proceed to create the scale. It can be used as a substitute for these four questions for reporting and analysis.

In a reliability analysis not shown, the reliability of the scale measuring the sales rep’s customer service is .79, also quite acceptable.

As a final point, what would happen if we included all seven questions in the reliability analysis simultaneously?

Click on the **Dialog Recall** tool , then on **Reliability Analysis**  
 Add the variables **repinfrm**, **repcourt**, and **reptime** in the **Items** list box  
 Click on **OK**

After the procedure runs, scroll to the bottom of the output. We see that, perhaps surprisingly, alpha is even higher, .895. Does this mean that we *could* place all seven questions in one scale, no matter the evidence from the factor analysis? The answer is a resounding maybe. First, recall our discussion of reliability and validity at the beginning of this chapter. Reliability is not somehow affected by validity, although it is a necessary condition for validity. Therefore, a high alpha for all seven items does not negate the factor results.

### Figure 19 Cronbach’s Alpha For All Seven Questions

Item-total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
REPANAL	13.4082	21.7758	.7664	.6778	.8731
REPENV	13.3673	20.3965	.7423	.6009	.8747
REPBUS	13.1633	20.9753	.7137	.6536	.8781
REPNEEDS	13.0816	20.9136	.7009	.6281	.8799
REPINFRM	13.4694	21.2312	.7321	.6256	.8758
REPCOURT	14.0000	23.7014	.6087	.5398	.8903
REPTIME	13.7347	22.1176	.6340	.5242	.8873
□					
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 )					
Reliability Coefficients      7 items					
Alpha =    .8954                      Standardized item alpha =    .8968					

But...it does suggest that the seven items could be placed in one scale, if that is theoretically justified and the scale was valid. Of course, we already know the items are measuring two broad constructs, so using them together on one scale would only be appropriate if these two constructs were complete indicators of a sales reps performance, as noted above. Then we would have a scale that would be quite useful. But if this isn't true, then there is really no compelling reason to use all seven together.

We finish this chapter by creating the four-item scale.

## Methods of Scale Creation

A scale is created by adding together the responses on the individual questions. There are three common methods to complete this task.

- 1) The more technically correct approach is to use scores that factor analysis can create for each respondent on each factor. These *factor scores* are calculated by multiplying the original variables by a set of weighting coefficients derived from the loadings. They are normed to have a mean of zero and a standard deviation of one. Although this method has its technical advantages, it has the disadvantage of using information from all the variables that have a non-zero loading, which defeats the purpose of creating a scale composed of a distinct set of questions.
- 2) A second approach is to pick only those items with the highest loadings on a factor—e.g., the four items on the first factor—and compute a new variable which is the sum or mean of that set of variables. This method keeps (if means are used), the new scores on the same scale as the original variables, which can make interpretation and presentation more straightforward.
- 3) A variant of this is to use the items with the highest loadings, but standardize them (create z scores) before adding them together. This technique should be used when the items have widely different variances.

## Creating the Scale

We will use the simpler method of creating a scale rather than employing factor scores. We add the responses and then divide by 4 to create a mean because it creates scores on the same scale (1 through 5) as the original responses.

The SPSS Compute procedure can be used to create the scale with the Mean function. First, though, we have another decision to make. As mentioned, a sizable proportion of respondents didn't answer all four





---

questions. Should the scale be created without their responses, or should we base the mean on either three or four valid responses? The Mean function has the capability to make this distinction and correctly calculate a mean for three instead of four items.

If we knew the data were missing at random, then we could substitute the mean value for a variable when a response is missing (just as with the factor analysis). But since we don't know the pattern of the missing data, we must take a different route. The most conservative approach is to use listwise deletion and only create a scale score when all four variables have valid values, but we lose over one-quarter of the respondents with this method. The alternative is to be more flexible and calculate a scale score if only three values are valid.

It is clear what this gains us: many more valid cases. What is the trade-off? First, is it even correct to use three rather than four items? After all, factor 1 has high loadings from all four questions. And, second, assuming this is acceptable, is the three-item scale reliable?

We can answer the second question with the output at hand. If we look at the output from Reliability (Figure 18), we observe that the alpha of all possible three-item scales is still quite satisfactory (look at the *Alpha If Item Deleted* column). So reliability won't be a problem.

To determine whether the scale would be valid, we need to do two things. First, we consider each three-item scale theoretically and ask ourselves whether it has content validity. And second, we could rerun the factor analysis four times, dropping one variable on each run. If we continue to get two factors, with the same items loading on factor 1, then we have strong support for using a three-item scale.

In practice, researchers rarely rerun the factor analysis, chiefly because factor analysis is an exploratory procedure with some imprecision. We are already using somewhat arbitrary cut-points to determine which variables load highly on a factor. And, we are using the less technically correct solution of using *only* variables with high loadings to create the scale score (by summing these variables), rather than using factor scores. Thus, most researchers simply assume that a scale missing one item is still valid.

A more telling criticism of this approach is that the scale will not be composed of the same items for all respondents. That is certainly true, and if this makes you uncomfortable, you need to compare the two versions of the scale—complete and incomplete—to determine which to use.

To remedy this problem, methodological variants are to 1) drop the item with the lowest loading (in this instance, repenv) or 2) drop the item with the most missing data (in this instance, repbus). Both these methods insure that the scale has the same items for all respondents.

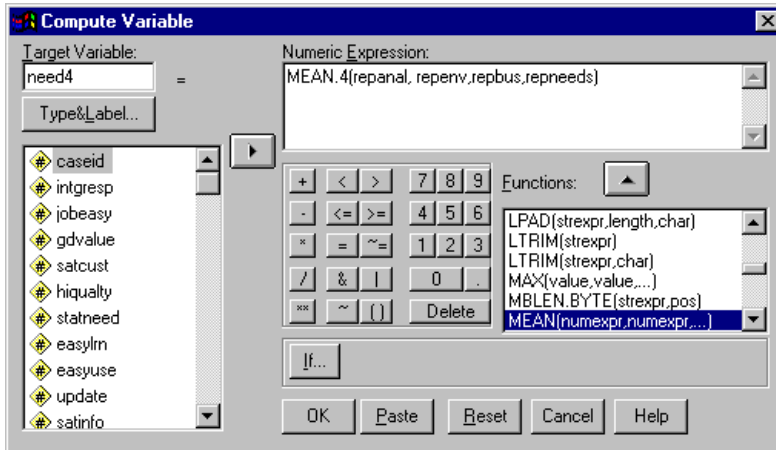
We don't have the time to try all these methods, so we create scales with only valid values and then with valid values for at least three of the variables. With this data, dropping the variable with the lowest loading only gains 5 cases. And the problem with dropping repbus, the variable with the most missing data, is that it has the highest loading on factor 1.

The SPSS Compute procedure is found under the Transformation menu

- Switch to the **Data Editor** window
- Click on **Transform...Compute**
- Click in the **Target Variable** box and type **need4** for the first scale
- Place the **Mean** function in the **Numeric Expression** box
- Add the variables **repanal**, **repenv**, **repbus**, and **reneeds** within the parentheses, with **commas** between them
- Type **.4** after the word **Mean** but before the left parenthesis


Your dialog box should look like Figure 20. By default, the mean function calculates a mean if any of the variables is valid for a case. This is not how we want to create the scale. The “.4” after the function name tells SPSS to calculate the mean only if all four values are valid.

**Figure 20 Completed Compute Variable Dialog Box**



Click on **OK** to create the new scale

The variable is added to the end of the Data Editor. Before examining it, let's create the second version of the scale.

- Click on the **Dialog Recall** tool , then on **Compute Variable** (not shown)
- Change the **Target Variable** name to **need3** (not shown)
- Click in the **Numeric Expression** text box and change Mean.4 to **Mean.3**
- Click **OK** to create the second scale

We now compare the two versions of the scale with the Descriptives procedure.

- Click on **Analyze...Descriptive Statistics...Descriptives**
- Place **need4** and **need3** in the **Variables** list box
- Click on **OK**

**Figure 21 Summary Statistics for Need4 and Need3**

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
NEED4	500	1.00	5.00	2.4575	.8763
NEED3	545	1.00	5.00	2.4839	.9024
Valid N (listwise)	500				

The mean and standard deviation for the two scales are essentially identical. This is very reassuring. There are 45 more valid cases for need3 since only three valid responses were required to create it. Need3 thus increases the sample size by almost 10% over need4, a substantial increase. Moreover, the reliability analysis showed that all three-item scales have alphas above .80, so a three-item scale is reliable.

Thus, if we are concerned about the amount of missing data and want to maximize the number of valid cases for future analyses, we might use need3 (if we are not bothered by the objections outlined above). Conversely, a more conservative approach is to use need4. Other considerations can affect the decision as well, such as how each scale relates to the important dependent variables or to demographic characteristics.

**Note**

There is no need to save the file changes at the end of this session.

**Summary**

We reviewed the concepts of reliability and validity in depth. We created an index, then used factor analysis and reliability analysis to determine what scales could be created from a set of questions. We concluded the by reviewing various methods to create the scale score.



