Scale Construction Notes

Jamie DeCoster

Department of Psychology
University of Alabama
348 Gordon Palmer Hall
Box 870348
Tuscaloosa, AL 35487-0348

Phone: (205) 348-4431
Fax: (205) 348-8648

June 5, 2000

These notes were prepared with the support of a grant from the Dutch Science Foundation to Gün Semin. I would like to thank Heather Claypool and Dick Heslin for comments made on earlier versions of these notes. If you wish to cite the contents of this document, the APA reference for them would be

DeCoster, J. (2005). *Scale Construction Notes.* Retrieved <month, day, and year you downloaded this file> from http://www.stat-help.com/notes.html

For future versions of these notes or help with data analysis visit
`http://www.stat-help.com`

Contents

Chapter 1

Introduction

- The purpose of scale construction is to design a questionnaire that provides a quantitative measurement of an abstract theoretical variable.
- Scale construction can be seen as a specialized area of survey design. Not all surveys are scales, however. Before trying to apply the procedures presented here to the construction of your questionnaire you should decide whether it really is a scale.

  The most obvious key is that a scale uses a moderately large number of items to measure a single construct. The items within a scale are typically *interchangeable*, meaning that the response to one item has exactly the same theoretical meaning as the response to any other item within the scale. Each item is designed to be a different way to "ask" about the same theoretical variable. Some questions may written so that more positive responses indicate less of the variable in question (see the note on reverse coding in section 2.3), but the expectation is that the magnitude of the correlations between items (whether positive or negative) should be relatively high throughout a scale.

  Additionally, after a participant completes a scale the responses are aggregated in some way, such as by taking the average, to produce the actual measurement. The specific response provided to any individual question is usually not of interest.
- Good scales possess both *validity* and *reliability*. A scale has validity if it properly represents the theoretical construct it is meant to measure. A scale has reliability if repeated measurements under the same circumstances tend to produce the same results. These two concepts are very important to scale construction and will be discussed in greater detail in chapter 4.
- Sometimes a single questionnaire contains items from several different scales mixed together. This is perfectly legitimate. In this case your items will not all be interchangeable - items making up different subscales will be slightly different. The items should, however, be interchangeable within each subscale. More information on working with these types of scales is provided in chapter 5.

Chapter 2

Creating Items

This chapter provides a set of guidelines for writing good scale items. These guidelines are designed to produce scales possessing high validity and reliability (discussed in chapter 4). While the chapter does present some specific "do's" and "don'ts," it is much better to keep the general principles in mind while writing questions than to simply obey the specific rules. Writing good scales is definitely an art rather than a science. Each section below introduces a basic guideline to be followed when writing items, provides a justification for it, and lists a few specific ways it can be followed.

2.1 Make the questions simple

- You want to make your items as simple and straightforward as possible so that your respondents are able to fill out the scale quickly and easily. Complicated scales can lead to misunderstandings and annoy respondents, reducing the likelihood that you obtain good data using the scale. Other researchers will also be more likely to use your scale if it can be completed in a short period of time.
- You should design your items so that they can be answered with very little instruction. The response format should be obvious even to someone who has not read your introductory paragraph. This is necessary because you will have a certain percentage of your respondents who will fail to read your instructions.
- Always avoid *double-barreled questions*, where the item actually combines two different questions into one. Consider the item "Do you think that the technical service department is prompt and helpful?" While some respondents may be able to quickly put down an opinion (if they universally like or dislike the technical service department), those who think that the service is prompt but unhelpful or the reverse would have difficulty answering the question. Any item that includes the word "and" should be closely examined to see if it is actually a double-barreled question.
- You should also avoid *nonmonotonic questions*, where people could provide the same answer to a question for two different reasons. Consider the item "Only people in the military should be allowed to personally own assault rifles." Someone could disagree with this statement either because they feel that non-military people should be allowed to own assault rifles or because they feel that no one should be allowed to own assault rifles.

2.2 Make the questions clear

- Unclear questions can cause respondents to interpret your items in different ways, reducing the likelihood that they will answer in ways that are related to the theoretical construct of interest. An item should be designed such that every respondent interprets it in the same way.
- You should avoid using any vague words or phrases in your items. Consider the question "How do you feel about the organization you work for?" A respondent could interpret this as asking how happy

they are with their job, or how well they get along with their immediate superiors, or to what extent do they agree with the overarching political agenda of the company.

- The level of language used in the questions should be appropriate to the intended target audience of the scale. Scales being developed for use on the common population should use simple language and avoid the use of all technical terms and jargon. Scales intended for more specialized audiences can use more technical language, but you should still be sure that every person that might reasonably be measured using the scale would understand a term before using it in the scale. When in doubt use simpler language.

- Be wary about asking for respondents' opinions about social groups or political structures. These questions assume that the respondent has a single, uniform opinion about the entire entity, which may or may not be true. For example, the question "How do you feel about your parents?" would be difficult to answer for someone who has positive feelings about one parent but negative feelings about another.

2.3 Avoid biased language

- If the language of your questions makes some responses seem more expected or desired than others you will be unlikely to get a true measurement of respondents' characteristics. Instead they will alter their answers in a way that makes them look better according to the language of the question. This will alter the actual meaning of the scale, moving it away from the theoretical construct of interest.

- You might consider telling your respondents that there are no right or wrong answers in the questionnaire instructions. This can reduce respondents' tendencies to provide what they believe are appropriate responses instead of their true answers.

- It is important to avoid using emotional words or phrases in your items. The mere inclusion of an emotionally-laden term in a question can bias respondents' answers. Your questions should be as neutral as possible to get the most accurate results.

- Emphasized text (underlined, italicized, or boldfaced) may sometimes bias responses to an item. For example, the emphasis in the question "To what extent do you think that psychological research has *practical* implications for society?" actually communicates the researcher's assumption that psychological research is not practical. Responses to this question would likely be more negative than to responses to a similar question lacking the emphasis on "practical."

- When respondents complete questionnaires they may sometimes focus on providing responses on one end of the scale without paying a great deal of attention to the specific details of each question. Although you want do design your items so that they are theoretically interchangeable, you still want respondents to pay attention to each item so that each acts as a separate measurement of the theoretical construct.

  To encourage respondents to read each question more carefully you might try *reverse coding* a number of your items. To reverse code an item you just reword it so that lower responses indicate a larger amount of the abstract construct of interest. For example, you might rewrite the item "To what extent are you satisfied with your long distance provider?" as "To what extent do you feel that your long distance service could be improved?"

  It is important to make the language of your reverse coded items as straightforward as your normal items. In general it is not a good idea to reverse code an item simply by introducing a negative term. This usually leads to a confusing sentence, and respondents who are not careful may fail to miss the negative term and may incorrectly respond to the item. For example, the item "To what extent do you dislike your long distance provider?" could easily be misread as "To what extent do you *like* your long distance provider?" accidentally eliciting the opposite response.

  The inclusion of reverse coded items slightly complicates the way you score your scale. Before calculating an aggregate score you must first transform the responses on your reverse coded items so that

higher scores are associated with respondents possessing a greater amount of the underlying construct. Luckily there is a very simple way to do this. The score for reversed items should be calculated as (highest value + lowest value - selected response). So if your scale goes from 1 to 7 you would calculate the score for reverse coded items as (8 - selected response).

- One check you can perform to see if a given question is biased is to consider whether people are likely to use the full range of responses to an item. If you expect that responses would be clustered around one end of the response scale then it is probably not a good item. Either the wording of the question is biased or the question itself is not terribly informative. Try to rewrite the item in such a way that respondents would provide a broader range of answers.

2.4 Use a common structure

- As stated before, the purpose of a scale is to use a number of items to measure a single construct. This means that the items themselves should be designed in such a way so that the same response on two different items (after the appropriate reverse coding of responses) would indicate approximately the same level of the measured construct.

- Psychological scales almost universally use items in a Likert scale format. The important characteristic of a Likert scale is that it has respondents answer questions by picking a response on a numerical continuum. The following is an example of a Likert scale question.

```
How much do you like playing Monopoly?

     1          2          3          4          5          6          7
not at all                          somewhat                          very much
```

  Research has shown that Likert scales with seven response options are more reliable than equivalent items with greater or fewer options.

- You should try to use the same response scale for all your items. For example, you might make each item an attitude statement and then have responds answer using a scale ranging from "completely disagree" to "completely agree". This goes a long way to ensure that the numbers have the same meaning across your items. If this is not possible, then you should make sure that corresponding points on the response scales are approximately equivalent.

- The items should each be completely self-contained so that the actual order of presentation of the items could be randomized. For example, it is generally bad to have questions that refer to other items in the scale. This introduces dependence between questions (reducing validity) and makes administering the scale more difficult. If you are forced to use multiple response scales you would present them grouped by response scale, but within each group you should be able to randomize the question order.

Chapter 3

Data Collection and Entry

Once you have designed your scale the next step is to administer it to your target audience. Once the data is collected, it then must be entered into a computer data file so that it can be analyzed by your statistical software.

3.1 Administering the Scale

- You should apply standard experimental procedures when testing your scale. Specifically, you should try to exert as much control as possible over the test environment. Each respondent should complete the scale under relatively similar conditions. Environmental differences can add random error to your measurements, reducing the measured reliability of your scale. If you are collecting data for a pretest of the scale (see chapter 6 for a discussion of pretest versus test data) you may choose to use more relaxed procedures to make data collection easier.
- In your initial tests you will probably want to determine the relationships between your scale and other theoretically related scales. Since scales are generally designed to be quick and easy to use you can probably do all of this in a single experimental session.
- If you designed your items so that the presentation order can be randomized you may want to test several different versions of your scale at the same time. This way you can test for effects of question order, and, assuming there are none, can state that the results of your scale are independent of order.
- Most scales are fairly short so you might be able to make efficient use of research participants by conducting other experiments in the same session that they complete the scale. Before deciding to do this you should always consider any possible effects these other tasks might have on the way respondents fill out the scale.

3.2 Entering the Data

- Perhaps the best strategy to remove data entry errors is to have the questionnaires entered into the computer twice, in two different files. This is referred to as *double entry*. Most modern word processors have a "Document Compare" function which allows you to locate differences between two files. Data that has been entered incorrectly in one file should show up as mismatching the other file (since it is highly unlikely that the two files would contain the same typographical error on the same item). It is easy to get even 99% accuracy using this method. You can either have two different people type in the data or you can have the same person type it in twice - it actually makes little difference.
- You should have participants either write or circle a number when responding to your items. You will want your data in numerical form for analysis, and it will be much easier to enter if your typist can directly read the answers. Poorer options would be to have responses labeled with letters, or to have respondents make a mark along a continuum to indicate their answer. While the latter may appear

to provide participants with a more precise way of responding to the item, research has shown that it does not increase validity or reliability, and it makes data entry much more difficult.

- If your study requires a large amount of data entry (for example, if you are including your scale in a package of questionnaires) you should include data-entry references right on the printed surveys. For example, if you are using a spreadsheet for data entry you might write what variable the response corresponds to in the right-hand column of the survey.

3.3 Using Computer Software

- There are a number of programs currently available that allow for the computerized administration of surveys. The benefits of doing this are a reduced cost in survey administration (no photocopies) and the complete removal of the data entry procedure (as the data are already on the computer). The downsides are that the administrator must learn how to use the particular software involved, and the number of respondents who may complete the survey at the same time is limited by the number of computers that are available.

  It is usually easiest to put your survey on a computer if the questions have a similar structure and use the same response scale. While the exact details of what is involved in conducting computer surveys varies with each package, most of today's software is very straightforward and easy to use.

- If you decide to use paper surveys, there are a number of different types of software that you can use to enter your data. The three most common options are to use either a word processor (like WordPerfect or Microsoft Word), a spreadsheet (like Lotus 123 or Microsoft Excel), or a database (like Paradox or Microsoft Access).

  The best option is probably to use a spreadsheet program. Spreadsheets are generally much easier to work with for data entry than either of the other two options. Additionally, most statistical programs have procedures that allow you to directly translate spreadsheet files to data files. If you use a word processor you will probably have to save the data as a text file and then go through a possibly lengthy data import procedure. If you use a database you will likely have save the data as a spreadsheet file anyway for it to be read by your statistical software.

- If you use a spreadsheet or database but still wish to use double entry, you should export your data to a word processor and use the latter program to check for inconsistencies. This will typically be faster than using the spreadsheet or database compare functions, which are usually much more complicated.

- If you decide to use a word processor for data entry you are probably best off making your entries space delimited (where each response is separated from the surrounding ones by spaces) rather than column delimited (where each response is placed in a specific column of the file). The data import procedures for space-delimited files are usually less tedious than those for column-delimited files.

Chapter 4

Validity and Reliability

4.1 Overview

- Validity and reliability are independent of each other. Validity is often thought of as the "accuracy" of the scale while reliability is its "precision." Scales that lack validity have systematic biases to them, while those that lack reliability have large random errors associated with their measurement. Obviously, you want your scale to be as valid and reliable as possible.
- For a better understanding of the different contributions of validity and reliability we can think about scales as marksmen trying to shoot a bullseye. Consider the illustrations in figure 4.1.
    - In figure 4.1a we see a demonstration of a very poor marksman. The shots are aimed off-center and are spread over a large area. This is analagous to a scale that lacks both validity and reliability. Such a scale does not accurately measure the psychological construct of interest, and it also will be difficult to relate it to other measures because of the large measurement error.
    - The marksman in figure 4.1b is very precise: all of the shots fall fairly close to each other. Unfortunately they are *precisely* aimed off center. This is analagous to a scale that has good reliability but lacks validity. Such a scale will generate consistent results, but measures something other than the theoretical construct it was designed to represent.
    - A different problem appears to plague the marksman of figure 4.1c. While the shots on average appear to be aimed properly at the bullseye, this person appears to lack control causing the shots to be spread across a large part of the target. This is analagous to a scale that has validity but lacks reliability. While such a scale does accurately represent the target variable, the lack of measurement consistency can make it very difficult to use responses to the scale.
    - The final illustration (figure 4.1d) shows the results of the ideal marksman. The shots are both highly accurate (centered on the bullseye) and very precise (being close together). Similarly, the ideal scale both has strong validity and high reliability. To draw accurate theoretical conclusions from experiments using a scale, the scale must properly represent the psychological construct of interest. Additionally, it is much easier to find significant relationships between variables when they are obtained using reliable measurements.
- Both validity and reliability are continuous measurements (rather than dichotomous). For example, a scale can be "mostly" valid or "somewhat" reliable.

-Insert Figure 4.1 here-

Figure 4.1: Targets shot by different marksmen.

- Errors in measurement are always either the result of systematic biasing of your instrument or random error. Validity is a reference to the extent that your measuring instrument is biased, while reliability is a reference to the extent that your measuring instrument introduces random error to its results. If you know a scale's validity and reliability you have everything you need to judge how well the scale meets the purpose of its design.

4.2 Validity

- Researchers worry about many different types of validity. Cook and Campbell (1979) provide four major divisions of experimental validity: Statistical conclusion validity, internal validity, construct validity, and external validity. The validity that we are most concerned about when creating a scale is *construct validity*. Construct validity is the extent to which the measurements taken in a study properly represent the underlying theoretical constructs.

- When we attempt to validate a scale we try to demonstrate that our theoretical interpretation of the responses to the scale is correct. Validity therefore measures the match between a variable representing a "true" measure of the construct and the scale responses. A scale by itself is therefore neither valid or invalid. The question of validity comes in only when we attempt to relate the scale to a particular theoretical construct. A scale could be valid for one purpose but invalid for another.

- Under the most ideal conditions there is an objectively correct way to measure the underlying construct your scale was designed to represent. In this case you can demonstrate that your scale has *criterion validity* by showing that the scale is related to the correct measure (the criterion). When such a criterion exists the only thing that matters regarding validity is the relationship between your scale and the criterion. Demonstrating that the two are related is sufficient to conclude that the scale is valid. Additionally, no other demonstration can validate your scale if it is not related to the criterion.

- Most of the time, however, there are no criteria from which to obtain an objective measurement of the construct underlying your scale. Variables measured by scales are usually too abstract to be measured objectively, which is why the scale is being developed in the first place. In this circumstance there is no single procedure that you can perform to measure validity. Rather, you must attempt to build an argument for a particular interpretation of the scale by demonstrating that the measurements are consistent with the theoretical variable that is supposed to be motivating the responses.

- One method is to argue that the scale possesses *face validity*. A scale has face validity if the items composing the scale are logically related to the underlying construct. In essence, face validity asks whether the scale "looks" appropriate. Face validity is usually a strictly qualitative judgment and is most convincing if supported with other more objective data.

- For scales that lack objective criteria it is probably most important to demonstrate that the scale has *convergent validity*. To do this you show that the responses to your scale are related to other measurements that are supposed to be affected by the same variable. For example, suppose that you are building a scale to measure social leadership. It seems reasonable that people which strong social leadership tendencies would also be extroverts. To help validate the scale you might therefore demonstrate that responses to your social leadership scale correlate with responses to an introversion-extroversion scale.

  You can (and should) assess convergent validity in a number of different ways. Each time you demonstrate that the scale acts in a way consistent with the underlying construct you make a more convincing argument that the scale provides an accurate representation of that construct.

- When assessing convergent validity it is often useful to simultaneously assess *divergent validity*. To demonstrate divergent validity you show that your scale is *not* related to measurements that are supposed to represent different variables. This is most useful if the diverging measurement is somehow similar, but for important theoretical reasons should not be related to responses to your scale. This can help demonstrate that your scale is measuring a new concept, rather than simply duplicating an

existing scale. For example, again consider the scale measuring social leadership presented above. Let us suppose that social leadership is not supposed to be related to autocratic leadership. We might therefore attempt to show that scores on the social leadership scale are uncorrelated with scores on an autocratic leadership scale.

Divergent validity is best assessed at the same time as convergent validity. There is always a statistical problem with demonstrating the lack of a relationship between two measurements, since it is unclear whether the relationship does not exist or if your study simply lacked sufficient power to detect it. However, if you conduct an investigation and are able to demonstrate some significant relationships (i.e., the ones designed to show convergent validity), it makes the argument that the nonsignificant findings in your assessment of divergent validity are due to faults in your experiment much less viable.

- At the same time that you demonstrate your scale's validity you may wish also to demonstrate the scale's *unique utility*. To do this you demonstrate that your scale does something above and beyond similar measures that already exist. This typically involves showing that your scale can explain unique portions of the variance in important variables. While this is not really an issue of validity, it is a factor that can cause people to use or not use the scale.
- Although validity and reliability are defined to be independent concepts, they are related in an important way. It is very difficult to determine the validity of a highly unreliable scale. Establishing criterion, convergent, and divergent validity all involve showing statistically significant relationships between the scale and other measures. If your scale is unreliable such relationships will be hard to demonstrate. You may have a valid scale in those cases, but you will not be able to show it.
- Keep in mind that validity measures how successfully your scale matches onto the theory that you propose. If you fail to validate the scale, it does not necessarily mean that there is something wrong with the scale. It can also indicate that there is a problem with the theory upon which you are basing your validation.

4.3 Reliability

- Determining the reliability of a scale is somewhat different from determining the validity of a scale. Unlike validity, reliability is a precisely defined mathematical concept. Reliability is measured on a scale of 0 to 1, where higher values represent greater reliability.
- Ideally, the measurements that we would take with our scale would always replicate perfectly. However, in the real world there are a number of external random factors that can affect the way that respondents provide answers to the scale. A particular measurement taken with the scale is therefore composed of two factors: the theoretical "true score" of the scale and the variation caused by random factors. This relationship is summarized in the equation

$$\mathrm{M} = \mathrm{T} + e, \tag{4.1}$$

where M is the actual scale measurement, T is the theoretical true score, and e is random error. The random error could be either a positive or negative value.

The true score represents the average score that would be obtained if a person were measured an infinite number of times. It is important to note that it has *nothing* to do with the theoretical construct behind the scale. It is simply a measure of whatever systematic component happens to exist in the measurements. The question of whether this systematic component matches onto the proposed theoretical construct is solely a question of validity, and is irrelevant when determining reliability.

- The reliability coefficient ρ is defined as

$$\rho = \frac{\sigma^2\{\mathrm{T}\}}{\sigma^2\{\mathrm{M}\}}, \tag{4.2}$$

where $\sigma^2\{\mathrm{T}\}$ is the variance of the scale's true score and $\sigma^2\{\mathrm{M}\}$ is the variance of the actual observed responses. It therefore represents the proportion of the variability in the observed scores that can be attributed to systematic elements of the scale.

- Unfortunately we cannot use formula 4.2 to calculate reliability because we can never measure $\sigma^2\{\mathrm{T}\}$. However, using this equation statisticians have derived several methods of calculating the reliability coefficient using real data.

- One way to calculate reliability is to correlate the scores on *parallel measurements* of the scale. Two measurements are defined as parallel if they are distinct (are based on different data) but equivalent (such that you expect responses to the two measurements to have the same true score). The two measurements must be performed on the same (or matched) respondents so that the correlation can be performed. There are a number of different ways to measure reliability using parallel measurements. Some examples are

    - *Test-Retest method.* In this method you have respondents complete the scale at two different points in time. The reliability of the scale can then be estimated by the correlation between the two scores. The accuracy of this method rests on the assumption that the participants are fundamentally the same (i.e., possess the same true score on your scale) during your two test periods. One common problem is that completing the scale the first time can change the way that respondents complete the scale the second time. If they remember any of their specific responses, for example, it could artificially inflate your reliability estimate. When using this method you should present evidence that this is not an issue.
    - *Alternate Forms method.* This method, also referred to as *parallel forms*, is basically the same as the Test-Retest method, but with the use of different versions of the scale during each session to reduce the likelihood that the first application of the scale influences responses to the second. The reliability of the scale can then be estimated by the correlation between the two scores. When using alternate forms you should show both that the application of the first scale did not affect responses to the second and that the two versions of your scale are essentially the same. The use of this method is generally preferred to the Test-Retest method.
    - *Split-Halves method.* One difficulty with both the Test-Retest method and the Alternate Forms method is that the scale responses must be collected at two different points in time. This requires more work and introduces the possibility that some natural event might change the actual true score between the two applications of the scale. In the Split-Halves method you only have respondents fill out your scale one time. You then divide your scale items into two sections (such as the even-numbered items and the odd-numbered items) and calculate scores for each half. You then determine the correlation between these two scores. Unlike the other methods, this correlation does *not* estimate your scales reliability. Instead you get your estimate using the formula

      $$\hat{\rho} = \frac{2r}{1+r}, \tag{4.3}$$

      where $\hat{\rho}$ is the reliability estimate and r is the correlation that you obtain.

      Note that if you split your scale in different ways you will obtain different reliability estimates. Assuming that there are no confounding variables, they should all be centered on the true reliability. In general it is best *not* to use the first half and second half of the questionnaire since respondents may become tired as they work through the scale. This would mean that you would expect greater variability in the score from the second half than in the score from the first half. In this case your two measurements are not actually parallel, making your reliability estimate invalid. A more acceptable method would be to divide your scale into sections of odd-numbered and even-numbered items.

- Another way to calculate reliability is to use a measure of *internal consistency*. The most popular of these reliability estimates is Cronbach's alpha. Cronbach's alpha can be obtained using the equation

  $$\alpha = \frac{N\bar{r}}{1+\bar{r}(N-1)}, \tag{4.4}$$

  where α is Cronbach's alpha, N is the number of items in the scale, and $\bar{r}$ is the mean interitem correlation. From the equation we can see that α increases both with increasing $\bar{r}$ as well with increasing N.

Calculating Cronbach's alpha is the most commonly used procedure to estimate reliability. It is highly accurate and has the advantage of only requiring a single application of the scale. The only real disadvantage is that it is difficult to calculate by hand, requiring you to calculate the correlation between every single pair of items in your scale. This is rarely an issue, however, as all the major statistical packages have procedures that will calculate α automatically.

- The reliability of a scale is heavily dependent on the number of items composing the scale. Even using items with poor internal consistency you can get a reliable scale if your scale is long enough. For example, 10 items that have an average interitem correlation of only .2 will produce a scale with a reliability of .714. However, the benefit of adding additional items decreases as the scale grows larger, and mostly disappears after 20 items.

  One consequence of this is that adding extra items to a scale will generally increase the scale's reliability, even if the new items are not particularly good. An item will have to significantly lower the average interitem correlation for it to have a negative impact on reliability.

- Like all statistical estimates, the confidence you can put into your reliability estimate increases as you test your scale using more respondents. You should therefore use at least 20 participants when calculating reliability. Obtaining more data won't hurt, but will not strongly impact the stability of your findings.

- Reliability has specific implications for the utility of your scale. Specifically, the most that responses to your scale can correlate with any other variable is equal to $\sqrt{\rho}$. The variability in your measure will prevent anything higher. Therefore, the higher the reliability of your scale, the easier it is to obtain significant findings. This is probably what you should think about when you want to determine if your scale has a high enough reliability.

- It should be noted that low reliability does *not* call results obtained using a scale into question. Low reliability only hurts your chances of finding significant results. It cannot cause you to obtain false significance. If anything, finding significant findings with an unreliable scale indicates that you have discovered a particularly strong effect, since it was able to overcome the hindrances of your unreliable scale. In this way using a scale with low reliability is analogous to conducting an experiment with a low number of participants.

4.4 Using Statistical Software

- While you will likely use statistical software in your attempts to establish the validity of your scale, these analyses will not call upon any specific procedures. Rather, you will use whatever procedures are most appropriate to test for each particular relationship you wish to demonstrate. So you might use correlations, regression, or ANOVA, depending on the type of data that you are working with.

- If you want to calculate the reliability of your scale using parallel measurements, any statistical program can be used to determine the correlation between the two scale outcomes. You will want your data set structured so that each case represents a respondent. You will then have two variables representing the two scale measurements for that respondent. You would then be interested in the correlation between these two variables.

- If you want to calculate the reliability of your scale using Cronbach's alpha, both SAS and SPSS have procedures that will do this automatically. With either program you will want your data set structured so that each case represents a respondent. You will then have a variable to hold the respondent's answer to each item on the scale.

  - In SAS Cronbach's alpha is calculated using the **alpha** option on **proc corr**. The following example code will calculate the reliability of five scale items contained in the variables v1 through v5.

    ```
    proc corr alpha;
      var v1 v2 v3 v4 v5;
    ```

At the top of the output SAS will print out the value of Cronbach's alpha for both the raw and standardized variables. The value that you would want to report is the alpha for raw variables. Below this SAS provides you with information about individual items in your scale. You will see the item-total score correlation and what the alpha would be if you removed the item from your scale. If removing the item from your scale would actually increase reliability you should probably either rewrite the item or drop it from your scale (see chapter 5). Finally SAS provides you with the correlation matrix of your items.

- Procedures for Reliability analysis can be accessed in SPSS versions 6 and 7 by selecting **Statistics → Scale → Reliability Analysis**. In later versions you select **Analysis → Scale → Reliability Analysis**. After that you will be taken to a variable selection screen. You must select all of the variables that represent items in your scale. After that, clicking **OK** will cause SPSS to calculate the reliability of those items.

  By default SPSS uses Cronbach's alpha, and does not produce any statistics other than the actual reliability estimate. If you wish to use another method to estimate reliability, there is a drop-down menu on the lower left part of the variable selection screen that will allow you to determine the procedure used by SPSS. If you want SPSS to produce information about how each item impacts on the scale's reliability you should click **Statistics...** in the variable selection screen and then check the **Statistics if item deleted** box. There are other checkboxes here that can provide you with additional output, such as the inter-item correlations. Once you have made your selections you click **Continue** to return to the variable selection screen. As before, to receive your reliability analysis from the variable selection screen you click **OK**.

Chapter 5

Measuring More than One Variable with a Scale

5.1 Subscales

- There are two basic situations where you might calculate more than one variable from the items in your scale. The first situation is when you believe that all of your items are related to the same abstract variable, but you expect that the answers to your items will organized in clusters. In this case you are testing a single scale possessing *subscales*. Statistically, all the items in your scale should be correlated, but you expect that the items will be organized in groups with even higher internal correlations.

  For example, you might create a questionnaire designed to measure political liberalism and conservatism. There are a number of different issues that distinguish liberals from conservatives, such as funding for the military, environmental regulations, and socialized medicine. You might create a group of items to determine respondents' opinions on each of these issues. You would obtain your liberalism/conservatism measurement by calculating the average score of all of the items in your scale. However, you might also want to calculate subscale scores for sets of items relating to the same political issue.

- In addition to calculating an overall score for your scale you must also calculate scores for each of your subscales. To do this you simply take the average of the items belonging to each subscale. You also can perform separate validity and reliability analyses for each of the subscales, if you suspect that they might individually be useful as predictors. To justify the use of calculating subscale scores you should demonstrate that there are circumstances when the use of particular subscales is preferable to the use of the overall scale. If the overall scale performs as well or better than each of the subscales then there is little point in calculating subscale scores.

5.2 Concurrent Scales

- In the second situation the items in your questionnaire are actually *not* all related to the same underlying abstract variable. Rather, different groups of items are related to different variables. In this case what you are doing is testing *concurrent scales*. Statistically, you think that items in the same group will be highly correlated but will not necessarily correlate with items in other groups.

  For example, you might construct a scale to measure a number of different personality traits, such as emotionality, aggressiveness, and extroversion. You would expect that items related to the same trait would produce similar responses, but you do not necessarily expect that there will be any consistency across the entire questionnaire. Therefore you would be interested in calculating scores for each trait, but you would not be interested in any type of overall score.

- When you test concurrent scales you do *not* want to calculate a score averaging over all the items. You do not expect that the average of all of your scales has any meaning, so there is no reason to calculate this score. You simply calculate separate scores for each scale in the survey, as if it were administered by itself. Additionally, the reliability and validity of each scale should be assessed independently.

5.3 The Proper Use of Factor Analysis

- It is fairly common to see factor analyses included in reports of surveys containing either subscales or concurrent scales. However, it is quite rare that the proper inferences are drawn from the analysis. This section will attempt to explain the most common error, and describe how factor analysis may be appropriately used to examine scale validity.

- To appreciate this issue it is key to understand that there are two different types of factor analysis. *Exploratory factor analysis* attempts to discover the nature of the constructs influencing a set of responses. *Confirmatory factor analysis*, on the other hand, tests whether a specified set of constructs is influencing responses in a predicted way. For more information you can examine the paper "An Overview of Factor Analysis," posted on the *Statistics Explained* website.

- Regarding scale construction, you will most often see researchers presenting exploratory factor analyses of their questionnaire in an attempt to justify the concurrent scales or subscales that they have defined. If the factors produced by the analysis appear to match the item groups that they have proposed, then they will claim that this is evidence for the validity of their theoretical interpretation of the scale.

- In reality, this provides *no* actual evidence for validity. First off, exploratory factor analysis is not designed to test theoretical predictions. That may only be performed using confirmatory factor analysis. However, it is much more difficult to perform a confirmatory factor analysis, so researchers will often instead choose to perform an exploratory factor analysis and simply discuss it as if it provides a statistical test. This is scientifically invalid, and is a misinterpretation of the calculated statistics.

  Secondly, the information provided by a factor analysis (of either type) is generally not what we are interested in when we want to test the validity of a scale. All that factor analysis can tell us is whether the responses to the survey are organized into particular clusters or not. It cannot tell us anything about *why* the answers are clustered in the way that they are. Since authors have complete control over the way that the items are written and presented, it is quite possible that items they perceive as theoretically related are written using similar language and with similar tone. Sometimes theoretically grouped items are even presented together under a label, which would influence respondents even more. Theoretical differences between items are almost always confounded with other characteristics of the question, so the results of a factor analysis typically do not say much about the variables underlying scale responses.

- The only time it is truly appropriate to use a factor analysis in the examination of scale validity is when the items in the questionnaire can be thought of as representative of some theoretically meaningful population of items. In this case you have reason to believe that the factors appearing in your scale would also be present in the larger population. If you can demonstrate that the factors are based on theoretically important variables you would then have evidence for claiming that these variables underlie responses to the questionnaire.

  If you want to claim that the results of the analysis provide support for the influence of specific theoretical constructs on the responses you must use confirmatory factor analysis. This allows you to specify a model, indicating how you think the responses will be grouped. The analysis will then provide a statistical test of how well the model actually fits the observed data. Exploratory factor analysis, on the other hand, can only be appropriately used as a description of what you found, not as evidence supporting a particular prediction.

Chapter 6

The Role of Revision

- While researchers sometimes have a fixed set of items that they know must compose their scale, most of the time scale items are heavily edited to provide the greatest validity and reliability. A scale will typically be tested and revised three or four times prior to professional presentation.
- This discussion will discriminate between "pretests" and "tests" of a scale. Pretests are when you collect data purely so that you can know how to revise your scale. Tests are when you collect data that you expect will be used to establish the validity and the reliability of your scale in some professional forum. The distinction is important because you typically take a number of liberties when collecting and analyzing pretest data that make the results invalid for formal reporting. You must always perform a test of the final version of the scale to justify its use . Prior to this there may be several additional tests and pretests.
- The typical procedure for developing a scale is

  1. Write out the original items in the scale. There should be a somewhat larger number of items than you would actually want in the final version of the scale, since you may decide to drop items during revision.
  2. Have another person read through the items and check them for clarity. This person does not necessarily need to be an expert in the topic being examined. It is more important that they have the confidence to point out items that are badly worded, overly complex, or simply difficult to understand. These comments can greatly help sharpen a scale.
  3. Administer the current version of the scale to a pretest sample. This should consist of at least 20 respondents, none of whom have seen the scale before.
  4. Conduct a preliminary reliability analysis. In this step you obtain the item-total score correlations and the reliability if each item were removed from the scale. Both of these can be obtained easily using either SAS or SPSS (see section 4.4).
  5. Conduct the primary reliability analysis. If you have a particular number of items that you want, simply select the appropriate number from those with the highest item-total score correlations. Otherwise, select all the items that would reduce the reliability if they were removed from the scale. Calculate the reliability of the scale including only the selected items.
  6. If the reliability is sufficiently high, proceed to step 7. The items you selected in step 5 compose the final version of your scale. Otherwise you should attempt to rewrite the items with the lowest item-total score correlations (making them more consistent with the abstract construct they are meant to represent) and return to step 3.
  7. Administer the final version of the scale to a test sample. In addition to the scale, you should have respondents complete whatever additional measures you will need for your validity analyses.
  8. Conduct final reliability analysis. Your reliability estimate should be approximately the same as the estimate calculated in step 5. If the reliability is significantly lower (such that you think the

power of your validity analyses may be endangered) you should treat the sample as a pretest and return to step 4.

9. Conduct validity analyses.

- As mentioned above, you do not have to be as careful with the experimental setup when administering pretests as you do when administering tests. Tests are used to demonstrate the features of your scale to a larger audience. Poor designs can cause you and other researchers to draw false conclusions about your scale and so must be avoided. Pretests, on the other hand, are only used to figure out how your items behave so that you know how to revise them for the next version of the scale. You will never report the findings of your pretests, so it is less important that they be conducted with great rigor. You might decide to violate some tenets of experimental methodology because it makes collecting the data easier. For example, it would be perfectly reasonable to pretest a work attitudes scale using people from a single company, but you would want to use a broader sample when testing the final version of the scale.

  You should keep in mind, however, that the purpose of experimental methodology is to improve the quality of inferences that can be made by examining a sample of data. If you have major violations of standard experimental procedure in the collection of your pretest data you may inadvertently revise your scale incorrectly.

- If the scale you are constructing contains subscales you will want to use more flexible criteria when deciding whether an item should be included or dropped from the scale. It is possible that an item might contribute strongly to its subscale but only weakly to the overall scale. You should make an explicit decision as to which is more important, or if they are equally important. You must, of course, include the same items in your tests of the scale and subscale reliabilities.

Chapter 7

Reporting a Scale

- You should always start your report with a discussion of the theoretical variables that you believe motivate responses to your scale. If the scale is the sole object of your presentation you should explain why the scale is needed and recommend ways that it can be used.
- Following this you should describe how you constructed the initial items, and why you feel that together they should represent the underlying theoretical construct. Providing a reasonable justification for how you created your initial items will increase the apparent face validity of the scale. You should report the number of items in the final version of the scale, as well as two or three example items so that the reader has an idea of what the exact content is like.
- You then present the statistical analyses that you performed to demonstrate the scale's validity and reliability. It is best to present the reliability scores first, since a low reliability can impact the interpretation of validity analyses.

  You should only present the validity and reliability analyses from your final test of the scale. The specifics of your revision process are not really important, and unnecessarily lengthen the statistical section of your report. You will therefore only report a small portion of the actual analyses performed during the construction of the scale.
- The validity and reliability analyses that you report should always be based on the full set of items administered at the time of the final test. It is statistically inappropriate to report validity or reliability analyses after dropping items that had low inter-item correlations in that particular test. Doing this artificially inflates the validity and reliability measurements. It takes advantage of idiosyncratic elements of that particular testing session that would not generalize to other applications of the scale. For those familiar with model-building procedures, it is like testing a regression model on the same data that you used to build the model.
- At the end you should include an appendix listing all of the items in the scale and any special procedures that might be required for scale administration and scoring. You should indicate what items (if any) are reverse coded. If you wish to report the item-total score correlations you would also do that in the appendix.
- The thing to remember is that the usefulness of a scale is as much dependent on the theoretical basis behind its development as it is on the specific procedures and analyses that you perform in its construction. This should be reflected in your report. You should resist the temptation to allow statistical calculations to dominate your discussion of the scale.

References

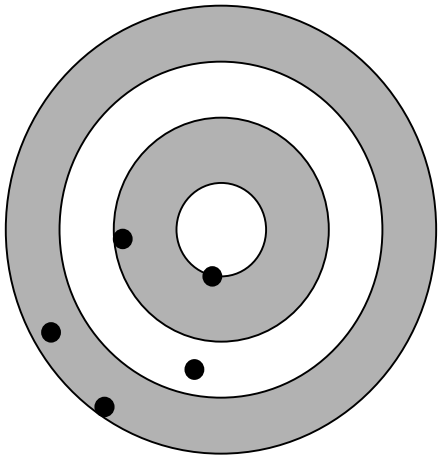
Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment.* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-017). Sage Publications: Beverly Hills, CA.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation. Design & Analysis Issues for Field Settings.* Houghton Mifflin Company: Boston, MA.

Rea, L. M., & Parker, R. A. (1997) *Designing and Conducting Survey Research: A Comprehensive Guide.* Jossey-Bass Publishers: San Francisco, CA.

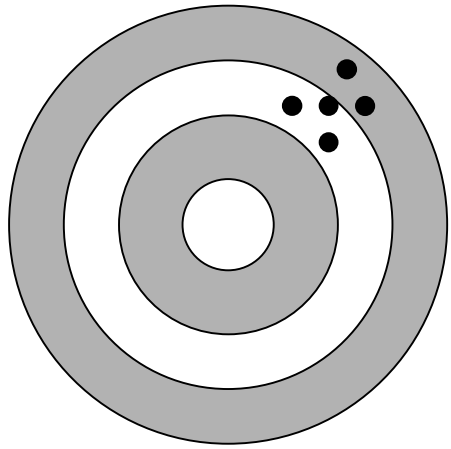
Rodeghier, M. (1996). *Surveys with confidence. A practical guide to survey research using SPSS.* SPSS Inc.: Chicago, IL.

a)



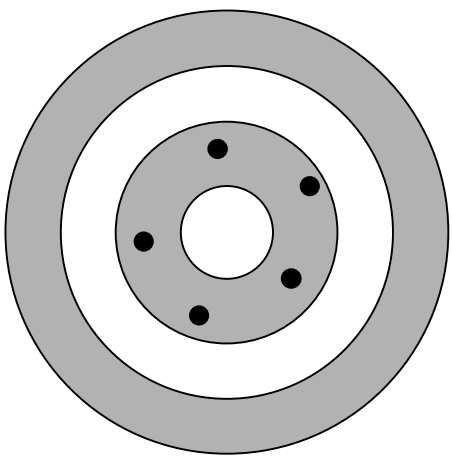
Low validity
Low reliability

b)



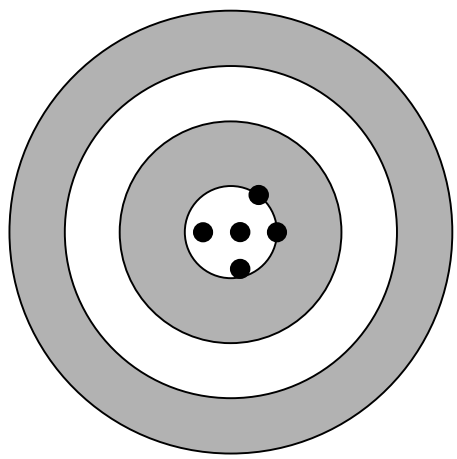
Low validity
High reliability

c)



High validity
Low reliability

d)



High validity
High reliability

Figure 1