CHAPTER 3

Designing and Scoring Responses

After working to have a solid conceptual definition of the construct, and then sweating over the operationalization of the construct by creating items, it is time to move on to the next step in the scale construction process by creating a response scale. This step calls for making decisions about what type of response you want from your test takers as well as the format of those responses. These issues may sound straightforward, but there are tradeoffs regardless of your choice.

Open-Ended Responses

The first major decision to make is whether open-ended responses or closed-ended responses are the most appropriate for the assessment situation. Open-ended items can be unstructured or structured. An example of a structured open-ended response would be to ask everyone in a class to write down their favorite prime-time television show on Sunday evenings. Then the frequency with which each response occurs can be counted, tabulated, and so forth. On the other hand, this type of response will not provide the test administrator with a lot of information about what motivates people to watch particular types of television shows. If a less structured, open-ended question is asked, such as, "Write down your favorite prime-time television show on Sunday evenings and *why* it is your favorite," then the administrator is likely to get some insight as to television-show-watching motives. However, wading through the responses and categorizing them into coherent groupings will take a lot of time.

Lots of detailed information is an advantage of responses to open-ended questions. Often open-ended questions are used in the early stages of theory development, or when an area of research provides conflicting findings. The open-ended approach can assist in clarifying what might be important to consider in including in theoretical frameworks. For example, there is considerable evidence that workplace stress is linked to cardiovascular problems (e.g., Cartwright & Cooper, 1997). However, the link is not consistent across individuals. Some researchers hypothesize that the link is moderated by such things as personality, gender, family history, social support, financial concerns, and so forth. Identifying these potentially moderating variables is greatly facilitated by responses of participants to open-ended questions such as, "What is most stressful about your work?" and "What types of things can you identify that alleviate workplace stress?"

The downside of open-ended questions is that they are very time-consuming to administer and interpret. The more open-ended questions asked, the more depth of information obtained, the more time it will take for respondents to complete the questions and the more time will be spent categorizing, analyzing, and trying to make sense of the responses. In fact, collecting and analyzing the qualitative responses from open-ended questions is a very detailed process that requires a strict methodology to ensure that researcher biases are minimized. Strategies to collect qualitative data include interviews, observation, the use of archival information such as notes and transcripts, and ethnography, to name but a few. The data collected provide rich and highly descriptive information that is very useful for understanding a phenomenon from a particular point of view in a particular context.

The qualitative data are content analyzed using a variety of techniques that include grounded theory, interrogative hypothesis testing, and case studies. Because the focus of this text is on quantitative measurement rather than on qualitative approaches, these techniques will not be pursued further. References that do an excellent job in describing how to collect and analyze qualitative data include Berg (1989), Creswell (1998), Denzin and Lincoln (2000), and Strauss and Corbin (1998).

Closed-Ended Questions

Closed-ended questions require a single response. The biggest drawback of closed-ended questions, in comparison to open-ended ones, is that the depth and richness of response is not captured. On the other hand, analyzing responses to these types of items is a relatively straightforward process. Frequencies of responses to closed-ended questions can be numerically coded and then depicted graphically. For example, bar or pie charts can be used to highlight different response categories for different groups. If relevant, the data might be shown in line graphs to show trends across time.

Closed-ended responses can also be analyzed and provide statistical evidence for making decisions (e.g., responses to a plebiscite on whether to widen a section of roadway in a city). A couple of examples of some simple analyses of responses to closed-ended questions will demonstrate how to use such data. Specifically, frequency analysis will be used to make some determinations about whether a particular type of television show is more or less common for the particular sample.

Table 3.1 Frequency Distribution for 30 Students and Their Television Preferences

Show Type	*Comedy*	*Drama*	*Life Situation*
Number (Percent) Preferring	17 (56.7%)	8 (26.7%)	5 (16.7%)

Example 1: Proportional Differences for a Single Variable.

Assume students in a class are asked to indicate which type of show they prefer most: situation comedies, dramas, or life situations. Of the 30 students, 17 prefer situation comedies, 8 prefer dramas, and the other 5 prefer life-situation shows. They indicate their show preferences as in Table 3.1.

To determine if the proportions are statistically different from that which would be expected in the population by chance alone (in this case, we would expect 33.3% of the cases to be in each category), a test of proportions is conducted. The formula for doing so is

$$(X - NP)/\sqrt{(N)\,(P)\,(1-P)}, \tag{3–1}$$

where X = number of responses in a category, N = total sample size, and P = expected proportion in the category.

So, in the example for situation comedies, $X = 17$, $N = 30$, and $P = 0.333$.

$$[17 - (30 \times 0.333)]/\sqrt{[(30)\,(0.333)\,(1 - 0.333)]},$$

$$(17 - 9.99)/\sqrt{(9.999)\,(0.667)},$$

$$7.01/2.58,$$

$$2.72.$$

The calculated value (2.72) is distributed approximately as a normal (z) distribution. Using the standard significance level of 0.05, the critical value to exceed to be considered statistically significant is 1.96 (using a two-tailed test). So, it can be seen that the ratio of 17 of 30 people watching situation comedies is statistically significant. Thus it can be concluded in this sample that a statistically significant proportion preferred comedy to the other two types of shows.

Example 2: Proportional Differences for Two Variables.

A somewhat more complicated question can also be asked of such data: Do men and women have similar or dissimilar tastes in Sunday prime-time television show viewing (using the same three categories of situation comedies, dramas, or life situations)? Assume there are 15 men and 15 women in the class and they indicate their show preferences as in Table 3.2.

Table 3.2 Frequency Distribution for 30 Students and Their Television Preferences by Gender

Gender	*Comedy*	*Drama*	*Life Situation*
Men	11	1	3
Women	4	9	2

The response pattern indicates that the men seem to have a higher than expected preference for comedy and the women seem to have a higher than expected preference for drama. An analysis of the data using the Pearson chi-square gives a value of 9.867. Box 3.1 shows the computation of the Pearson chi-square and Box 3.2 shows the SPSS cross-tabulation output that reports the Pearson chi-square.

The chi-square is calculated by examining the extent to which the expected cell frequencies deviate from the observed cell frequencies. If the deviations are significant, then it is concluded that the cell frequencies are dependent on the variables. In this example, the interest was in determining if show preference is dependent on gender.

Box 3.1 Computation of the Pearson chi-Square (χ^2) Using the Data From Table 3.2

The formula for calculating the Pearson chi-square statistic is

(3–2) χ^2 (Degrees of freedom of [(rows – 1) × (columns – 1)] = $\Sigma[(O_{ij} - E_{ij})^2/E_{ij}]$,

where O_{ij} = the observed frequencies for each cell and E_{ij} = the expected frequencies for each cell. That is, each cell's expected frequency is subtracted from the cell's observed frequency. These differences are squared, and then divided by the cell's expected frequency. Then the obtained values are summed across each cell.

The information from Table 3.2 is reproduced here as well as the row and column totals and row percentages (Table 3.3).

Table 3.3 Frequency Distribution for 30 Students and Their Television Preferences by Gender, Row and Column Totals, and Row Percentages

Gender	*Comedy*	*Drama*	*Life Situation*	*Row total*	*Row %*
Men	11	1	3	15	50%
Women	4	9	2	15	50%
Column total	15	10	5	30	100%

The first task is to generate the expected frequencies for each cell. To do this, we take the percentage of men and percentage of women and multiply each of these by the column totals to generate the expected cell frequencies. We see that men make up 50% of our sample. If the cells in the Men row were distributed as expected by chance alone, then we would have 50% of 15, or 7.5; 50% of 10, or 5; and 50% of 5, or 2.5 in the Men row. We would also have the same set of expected frequencies for the Women row.

These expected frequencies are noted in brackets along with the observed frequencies below (Table 3.4).

Table 3.4 Observed and Expected Frequencies (in Brackets) for a Distribution of 30 Students and Their Television Preferences by Gender

Gender	*Comedy*	*Drama*	*Life Situation*
Men	11 (7.5)	1 (5)	3 (2.5)
Women	4 (7.5)	9 (5)	2 (2.5)

Now, using our formula, we can calculate the Pearson χ^2:

$\chi^2(1 \times 2)$

$= \sum [(11 - 7.5)^2/7.5] + [(1 - 5)^2/5] + [(3 - 2.5)^2/2.5] + [(4 - 7.5)^2/7.5]$
$+[(9 - 5)^2/5] + [(2 - 2.5)^2/2.5]$

$= 1.63 + 3.2 + 0.1 + 1.63 + 3.2 + 0.1$

$\chi^2(2) = 9.86$

The computer output indicates that with two degrees of freedom [(no. of rows – 1) × (no. of columns – 1)], the obtained chi-square is statistically significant (0.007). This means that the pattern in the rows and columns is not what would be expected by chance alone (i.e., the rows and columns are not independent). The meaning of this finding would be interpreted by going back to the table to determine where the pattern seems to be unusual. In this case, 11/15 (73%) of the men preferred the comedy shows and 9/15 (60%) of the women preferred drama. These percentages are highly unlikely to occur simply by chance alone.

There is another statistic provided in the output that is called the *likelihood ratio chi-square statistic* (denoted as G^2). Like the Pearson chi-square, it has two degrees of freedom. However, the calculated value is 10.960 and its significance level is 0.004. The likelihood ratio is used more than any other statistic in multiway frequency table analyses. Although we have used only a simple two-way table here in

Box 3.2 SPSS Cross-Tabulation Output of the Data From Table 3.2

Chi-Square Tests				
	Value	*df*	*Sig. (two-sided)*	
Pearson Chi-Square	9.867	2	2	0.007
Likelihood Ratio	10.960	2	0.004	
N of Valid Cases	30			
Symmetric Measures				
	Value	*Approx. Sig.*		
Phi	0.573	0.007		
Cramer's V	0.573	0.007		
Contingency Coefficient	0.497	0.007		
N of Valid Cases	30			

this example, when tables become three-way, four-way, and so forth, the likelihood ratio is reported more often than the Pearson chi-square. We will not review the hand calculation for the likelihood ratio chi-square statistic in this text.

In addition to these chi-square statistics, there are summary measures of symmetry reported in the output. The ones of relevance are the *phi coefficient* (0.573), *Cramer's V* (0.573), and the *contingency coefficient* (0.497). The phi coefficient is simply a variation on the Pearson product-moment correlation coefficient. Pearson correlations are usually produced when correlating two continuous variables (like GPA and salary earnings as shown in Chapter 1). When the two variables are dichotomous, then the phi coefficient is calculated. A significant value indicates that there is a dependency in the data set; that is, if the value of one of the variables is known, there is a better than chance odds at guessing what the value of the other variable will be.

In our example, if it is known that the person is a man, it is likely that his preferred television show type on Sunday night at prime time is comedy. If it is known that the individual in question prefers drama television shows on Sunday nights at prime time, then it is likely that individual is a woman. Like any other correlation coefficient, the phi can take on positive or negative values. So, by noting the coefficient's sign and knowing how the nominal variables were coded, one is able to interpret how the relationships in the table are manifesting themselves.

Cramer's V is based on the calculated chi-square value and is a measure of the strength of a relationship between the variables. Because it is based on the chi-square value, it can only take on positive values. It is found via the following formula:

$$V = \sqrt{\chi^2/[(N)(n-1)]}, \tag{3–3}$$

where N = sample size and n = the number of rows or columns, whichever is smaller. In this case,

$$\begin{aligned} V &= \sqrt{9.867/[(30)(2-1)]}, \\ &= 0.573. \end{aligned}$$

The contingency coefficient is normalized slightly differently from phi and Cramer's V. Like the Cramer's V, it can only take on positive values (from 0–1).

These examples have demonstrated that when the responses to questions are categorized into frequencies, one can ask various research questions about the data. Simple one- and two-way frequencies have been reviewed here. However, it is possible to have more complicated designs where three or more variables are set up in frequency tables. These types of tables require multiway frequency analyses. Although they will not be reviewed here, a good source for information about these analyses is Rudas (1997).

Dichotomous Responses. Dichotomous responses are closed-ended questions that are most often coded with a 0 or a 1. They are frequently used when the item has a correct or incorrect response. For example, in this item, "Is the answer to 2 + 2 = 4 true or false?" the response is a dichotomous one. If the respondent answers "false," the code is 0 and if the respondent answers "true," the code is 1. This item response is a true dichotomy. With this particular item, the respondent has a 50/50 chance of getting the item correct just by guessing. Thus, care should then be taken in deciding if a dichotomous format with the potential for guessing is appropriate.

Another issue to consider is whether or not the item should actually require a dichotomous response. Consider the following item: "Do you feel happy today?" with response option of "yes" or "no." This dichotomy is somewhat limiting, and the response options to this question might be better asked on a continuum. This item response is thus called a false dichotomy. For example, the question can be rephrased to ask, "On a scale of 1–10, with 1 being extremely unhappy and 10 being extremely happy, how happy do you feel today?" While dichotomous responses force individuals to make a choice (i.e., yes or no) and this might suit the needs of the researcher or test administrator (e.g., quick to administer and easy to score), test takers may be reluctant to provide such responses. Because a dichotomous response format does not allow test takers the flexibility to show gradation in their attitudes, they may become frustrated and refuse to continue to complete the test or, if they do so, they may provide inaccurate information.

Dichotomous responses are also called for in responses to adjective checklists. A checklist presents a list of adjectives to respondents and they are asked to indicate

whether they think the adjective describes some stimulus (like oneself, a friend, coworker, spouse, etc.). For example, here is a list of adjectives and the respondent is directed to "check off the ones that characterize you":

1. quiet
2. sincere
3. happy
4. selfish
5. ambitious

Each response is then coded as a 1 if it is checked off and a 0 if it is not.

Multiple-Choice Tests. Multiple-choice tests are also called objective tests, and the items are scored as correct or incorrect. Thus, responses to these types of items are dichotomous. These types of tests are widely used and a number of issues arise when constructing, administering, and scoring them. These will be discussed next.

Distractors. One of the important aspects to multiple-choice test item creation is that the *distractors* (the options that are not correct) are just as important as the *target,* or correct response. Look at this item, "What is the sum of 15 + 365?" with the following four options provided: (a) 386, (b) 350, (c) 1478, and (d) 380. Which of the distractors is really not useful? The answer selected should be *c.* The first distractor would indicate that the person might be guessing, the second that he or she did an incorrect operation, and the fourth is correct. The third distractor is so outrageous that no one would likely pick that as a response. There are several guides to developing distractors, as there were with developing items. As before, these are based on Ghiselli, Campbell, and Zedek (1981) and Nunnally and Bernstein (1994).

1. Create distractors that are plausible, but not so plausible as to easily confuse the correct with the incorrect response.
2. Make all of the alternatives parallel in length and grammatical structure. If they are not, the correct response becomes more apparent.
3. Keep the alternatives short, putting as much of the information in the item stem as possible.
4. Don't write distractors that mean the same thing. The testwise student will know to eliminate them both as not correct.
5. Alternate the position of the correct answer within the distractors. Testwise students will figure out that, if the correct response is usually in the C position, then on an item to which they don't know the answer, a guess of C is better than chance.

6. Use the alternatives "all of the above" and "none of the above" as little as possible.
7. Make sure each alternative agrees with the stem. If it does not, then this is again a clue to the testwise student that the alternative is a distractor.

Analysis of distractor responding is usually done in a couple of ways. One is to see how many individuals selected each distractor. If a multiple-choice test has four options, then one would examine the percentage of the respondents choosing each distractor. If 5% or fewer respondents select the distractor, consider rewriting the distractor as it is not serving its intended purpose. Second, a sense of who is responding to the distractors can be obtained by carrying out chi-square analyses.

For example, suppose it is of interest to see if men or women are more likely to select a particular distractor to an item than would be expected by chance alone. A multiple-choice item is administered to 100 individuals (50 men and 50 women) and there are three response alternatives. Assume that response B is the correct response. The data collected can be shown in a table like that in Table 3.5. The computer output of the cross-tabulation analysis is shown there as well.

Table 3.5 Response Choice to an Item Cross-Classified by Gender and SPSS Cross-Tabulation Output

Gender	*A (distractor)*	*B (correct)*	*C (distractor)*
Men	10	30	10
Women	3	25	22

Chi-Square Tests			
	Value	*df*	*Sig. (two-sided)*
Pearson Chi-Square	8.724	2	0.013
Likelihood Ratio	9.044	2	0.011
N of Valid Cases	100		
Symmetric Measures			
	Value	*Approx. Sig.*	
Phi	0.295	0.013	
Cramer's V	0.295	0.013	
Contingency Coefficient	0.283	0.013	
N of Valid Cases	100		

Most of the participants (55%) answered the question correctly. The men were equally as likely to select Distractor A as Distractor C. The women, however, seemed much more likely to select Distractor C than Distractor A when they got the item incorrect. A 2×3 chi-square analysis run on this table indicates that, indeed, there is a significant dependency in the data; women are more likely to select Distractor C (22/50) than A (3/50). To determine if this is significant, the expected versus observed difference in proportions using Formula 3–1 can be applied.

In this example, for Distractor A, $X = 3$, $N = 25$ (25 women were incorrect), and $P = 0.50$ (it is expected by chance that 50% of the women selecting an incorrect answer would select Distractor A).

$$[3 - (25)(0.5)]/\sqrt{(25)(0.5)(1 - 0.5)},$$

$$(3 - 12.5)/\sqrt{(12.5)(0.5)},$$

$$-9.5/2.5,$$

$$-3.8.$$

Recall that the calculated value is distributed approximately as a normal (z) distribution and the critical value to exceed to be considered statistically significant is 1.96. The value of −3.8 indicates that Distractor A is significantly less likely to be selected than is Distractor C by the women.

Now that it has been demonstrated that Distractor A is selected disproportionately as a distractor, what can be done? The item writer must go back to the distractors and examine them closely. The task is to determine what it is about the content of the distractor that makes it more likely for women not getting the item correct to choose Distractor C. Test designers try to make the distractors consistent across demographic variables. This example used gender to check for demographic differences, but any type of grouping variable such as education level, race, high versus low test scorers, and so forth can be used to assess disproportionate distractor selection rates. Information based on gender or racial differences may point out potential problems with the item from a bias perspective.

Examining differences in distractor selection by those who did well on the test overall versus those who did more poorly provides information about whether distractors are useful for discriminating between better and worse performers, as well as whether or not the distractor might be confusing. Consider the situation where the best performers on the test overall all seem to get the answer to one item incorrect. Further assume that they all chose a particular distractor. On the other hand, the poorer performers did not select this distractor with any greater frequency than they selected another. This indicates that the item is tripping up the best students and the distractor might need to be rewritten.

Guessing. An issue in multiple-choice or true-false tests is that the respondent can guess the correct answer. For a true-false test, the respondent has a 50% chance of getting the answer correct without knowing the answer. For a four-option multiple-choice test, the respondent has a 25% chance of getting the

answer correct by guessing. Some tests factor a guessing penalty into computing the total score on a test to take this guessing component into account. The formula for doing so is

$$\text{Guessing corrected score} = C - [I/(n-1)], \tag{3–4}$$

where C = the number of correct responses, I = the number of incorrect responses, and n = the number of alternatives available for each item.

For example, assume a test has 100 multiple-choice items with five potential alternatives for each item. A test taker manages to complete 90 items. Of the 90 items, the test taker gets 70 correct.

$$\begin{aligned}\text{Corrected score} &= 70 - (20/4),\\ &= 70 - 5,\\ &= 65.\end{aligned}$$

Note that if the test taker had gotten none of the answered items incorrect, then the corrected score would not be lowered by a guessing correction. The more items the respondent gets incorrect, the more the person will be penalized for guessing on the correct responses. The assumption underlying this equation is that, if test takers get items incorrect, they are likely to be guessing.

It is extremely important as a test taker to know if a test has a penalty for guessing built into generating the total score. If there is no penalty for guessing and there is one minute left to complete the test but 20 questions left to answer, then the test taker should quickly choose any response to the rest of the items. However, if there is a penalty for guessing, test takers would not want to use this approach as it would be detrimental to their scores. The exception is if there are at least four alternatives and the test taker is able to narrow down the answer to two options; in this case, a guess provides a slight statistical advantage in the correction formula.

Speeded and Power Tests. Power tests assess individual differences without any effects of imposed time limits changing scores. Power tests often are made up of items that vary in their level of difficulty. Although pure power tests are not usual, most tests of achievement are designed so that 90% of the individuals taking the test can complete all of the items in a specified period of time (Nunnally & Bernstein, 1994). That is, most power tests of achievement have an arbitrary time frame (such as the length of a class period) as an administrative constraint. The power test designer should ensure that the test is long enough to cover the content domain, but not so long that there is not enough time available for most people to complete the test.

Pure speeded tests are composed of easy items where the variation in scores from individual to individual is based simply on how many correct items are completed. For example, a test in grade school where students are asked to complete as many multiplication facts as possible in two minutes is an example of a pure speeded test. Pure speeded tests are not useful unless the underlying construct being measured is one where speed is important (e.g., a typing task).

One issue in speeded tests that is somewhat different than that in a power test is the test length. Given the ease with which speeded test items can be constructed, it is usually not a problem to create new items if more are needed. The only way to get variance on the test scores for a group of test takers on a speeded test is to control the amount of time given to the test taker. Variability of test scores is a strong determinant of a test's reliability, so the time limit should be set to ensure maximum variance in the test scores. This can be determined empirically simply by having individuals complete as many items as possible in 1 minute, 2 minutes, 3 minutes, and so forth. The standard deviation can then be plotted against the time interval and a determination of the time limit when the highest variance in scores occurs is then selected. Figure 3.1 shows that the most variability for the items on the test in question occurs at 8 minutes. This, then, would be the optimal time limit on the test.

Some speeded tests often take into account the individual difference of age. It is a well-established empirical finding that as individuals age, their response times to stimuli slow (e.g., Birren & Schaie, 2001). Older individuals simply take longer to complete tasks than do younger individuals. When administering a speeded test, ensure that the test manual has addressed the issue of age. Usually this is done by adding a constant to scores of individuals within certain age bands. For example, individuals between 30–39 years of age taking the Wonderlic Personnel Test (1999) add one point to their raw score. Similar adjustments to raw scores are made for each decade up to 60 years of age and over.

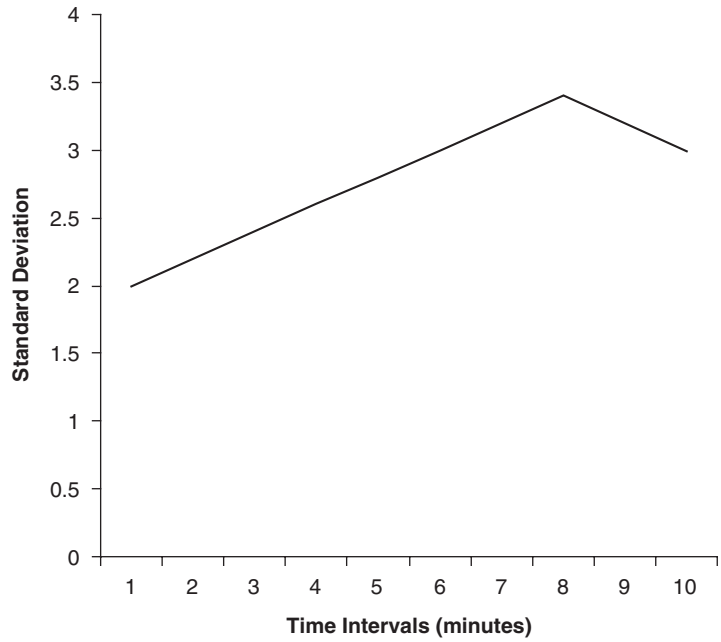

Figure 3.1 Speeded Test Standard Deviations

Omitted and Partial Credit. The terms *omitted* and *partial credit* are used in analyzing correct/incorrect test data and are particularly important in speeded tests. Omitted items are items that the respondent skips over. Sometimes it is appropriate to assign the omitted items a value of $1/A$, where A is equal to the number of alternatives. So if a respondent does not make a response to item 20 in a 25-item test with four alternatives, but completes items 1–19 and 21–25, the person would get a score of 0.25 on item 20. In effect, this gives the person a "guessed" value correct for the omitted items and improves the accuracy of score estimation (Lord, 1980). The formula for correcting for omitted items is

$$\text{Omitted corrected score} = 1/A \times O + \text{total}, \tag{3–5}$$

where A = the number of alternatives available, O = the number of omitted items, and total = the total number of items correct.

It may also be desirable to give partial credit for items where parts of the answer can be scored correctly and other parts incorrectly. It is important when using the partial credit approach that the difficulty of each of the parts of the question is known. For example, take an item such as the following:

$$\sqrt{25} + 2 = ?.$$

The answer to the first part ($\sqrt{25}$) requires more sophisticated math ability than does the second part ($+ 2$). Therefore, any partial credit strategy should be able to take into account differences in difficulty for the parts of the question. When the parts are of equal difficulty, then simply creating separate items out of each part would be appropriate. For example, in the following item:

$$(2 + 3) - 4 = ?,$$

the individual would get one point for solving the first part ($2 + 3$) correctly and one point for solving the second part ($5 - 4$) correctly. Test administrators should know in advance how omitted and partial credit are to be dealt with rather than trying to decide after the test takers have completed the items.

Continuous Responses

Up to this point, issues associated with responses that have only two primary options (dichotomous) have been reviewed. Many scales, however, are developed with responses that have more than two options for responding. One of the most popular of these types is the summated-rating scale based on the work of Likert.

Summated-Rating Scales. In the previous chapter, various ways to assess attitudinal items were presented. However, many of the procedures may have seemed long and tedious. Why can't a simple question such as, "What is your satisfaction with your coworkers?" be asked with the response being one of five options: very unsatisfied,

unsatisfied, neither satisfied nor unsatisfied, satisfied, and very satisfied? There are a couple of reasons why some care should be made in creating such items.

One assumption of items such as this is that the construct under investigation—in this case satisfaction with one's coworkers—can be placed into ordered categories where higher values can be inferred to mean higher satisfaction. Another assumption is that each of the response categories will have a normal distribution of responses around it in the population.

Recall that in the paired comparison approach example, 500 students were asked to make six paired comparisons between flavors of ice cream. Alternatively, those 500 students could have been asked to rate each of the four flavors of ice cream on a five-point scale, such as the following: "horrible, bad, okay, good, delicious." The problem with the latter approach is that there is no basis for determining the relative difference between "horrible" and "bad," nor is there any basis for determining whether the difference between "horrible" and "bad" is the same as that between "okay" and "good" (or for any other combination). Fortunately, a critical piece of empirical work was conducted such that confidence in the assumptions of these continuum-based response scales is now appropriate. The researcher who conducted the work was Rensis Likert.

In many ways, Likert revolutionized how attitudes are assessed and scaled. Prior to his work, the attitude scale assessment and scoring occurred as described in Chapter 2. In 1932, Likert published his method for scaling response categories that were separate from the items. He evaluated many attitude statements using five response categories: 1 = strongly approve, 2 = approve, 3 = undecided, 4 = disapprove, and 5 = strongly disapprove. As was shown in the previous chapter, these attitude items resulted in scale scores. What he found was that the simple categories 1, 2, 3, 4, and 5 that were labeled "strongly approve" to "strongly disapprove" correlated so highly with the more tediously determined scale scores that one could readily use the categories 1 through 5 rather than the item's scaled values.

This did two very important things. The first was to enable test developers to not be so dependent on the labor intensive generation of scale scores for each item. That is, it was no longer necessary to know "how much" stimuli was present in each item. The second was that the purpose of Likert's approach was not to scale items but to scale participants. One could obtain an assessment of an individual's strength of an attitude by simply summing across each of the response categories for each person.

So if respondents were presented with the 10 job characteristics that were introduced in Chapter 2, instead of having to respond "yes" or "no" as to whether each characterized their jobs, they could agree with the statement to a varying extent. They would do so by indicating their agreement level regarding how much each descriptor (such as "challenging") characterized their jobs on a scale anchored with 1 = strongly disagree, 2 = disagree, 3 = undecided, 4 = agree, and 5 = strongly agree.

It is critical to know the *valence* (negativeness or positiveness) of the item's content when calculating total scores on summated rating scales. Items with a negative valence should be *reverse coded* such that the response given is transposed: in the case of a five-point scale, a 1 would be changed to a 5, and a 2 would be changed to a 4,

while the 3 would remain the same. Recall that in Chapter 2, three of the items in the job description scale had negative scale scores: disgusting, underpaid, and revolting.

The job descriptors, a sample respondent's ratings of agreement, the item scale score values, and the ratings with reverse coding are shown in Table 3.6. Because of the reverse coding for the three items, the respondent's total score would be based on the last column of the table rather than the second column. In this case, the person's job satisfaction score would be 35. This 35 is an indication of the attitude of the respondent toward his or her job. This number can be compared to the scores of other job incumbents and used to make a decision such as whether or not the respondent should stay with the organization, or for other purposes. If a number of job incumbents complete the scale, the whole group's score on job attitude could be correlated with other variables, such as intentions to quit, organizational commitment, and so forth.

Likert designed his response approach to assess individual differences in attitudes. Scales today that use a five-point format like the one described are called Likert scales. Variations on the traditional Likert scale are called summated rating scales or Likert-type scales. These variations pose some issues that are explained next.

Variations and Issues With Likert Scales. Variations in the category descriptors seems to be the least problematic concern. While Likert used "strongly approve" to

Table 3.6 Likert Scaled Responses to Job Characteristics Items

Item	*1–5 Rating*	*Scale Score*	*Reverse Coded Responses*
disgusting	1[a]	–3.58	5
fun	2	1.51	2
underpaid	3[a]	–2.90	3
rewarding	4	1.66	4
delightful	2	2.56	2
challenging	4	0.70	4
enjoyable	3	0.89	3
revolting	1[a]	–3.49	5
interesting	3	1.08	3
meaningful	4	1.51	4
			$\Sigma = 35$

a. Negative valence items that are reverse coded.

"strongly disapprove" as his original descriptors, many scales use "strongly agree" to "strongly disagree" or "extremely difficult" to "extremely easy," and so forth. There is an assumption that the response categories of these variants are "equal interval" as Likert had demonstrated so many years ago. This assumption is not always correct. However, work by Bass, Cascio, and O'Connor (1974) as well as Spector (1976) suggest that most attitude surveys using Likert's general approach do have categories of approximately equal intervals.

The number of categories has also been a subject of much debate in the psychometric literature. Should there be 3, 5, 7, 9, or 26 categories, or does it matter at all? Symonds (1924) argued that 7 categories is the optimal number and Champney and Marshall (1939) indicated it is best to have 20 or so categories. Anderson (1991) found that for most rating tasks, 10 categories provides as much discrimination as is needed. Rather than pick a number of categories to be used based on empirical information, however, a more theoretical approach seems reasonable.

That is, what are the data going to be used for? Assume that a series of items with responses on a 1–5 scale with 1 = strongly disagree and 5 = strongly agree is posed. If categories 1 and 2 (strongly disagree and disagree) and categories 4 and 5 (agree and strongly agree) are going to be collapsed into single categories, then a three-point scale would have been sufficient. For example, in a performance appraisal rating task, a supervisor may be asked to rate the performance of an employee on a number of dimensions (punctuality, customer service, quality of work output, etc.). If the required information about the employee is "does not meet expectations," "meets expectations," and "exceeds expectations," then three categories are sufficient. Requiring the supervisor to rate the person on a 10-point scale might be stretching that person's capacity to make that fine a set of discriminations. On the other hand, if the sample of supervisors who will be the ultimate users of such a rating scale is used to having a 7-point scale for rating their employees, it may be difficult for them to adjust to a 3-point scale and they would not be comfortable using the 3-point scale.

So one should use the number of categories that provides as much discrimination as is reasonable to expect from the respondents combined with the number of categories to be used in any analysis and interpretation of results. It is most likely that a number between three and nine will be selected. Regardless of the number of categories selected, it is very helpful to have clear descriptors for each of the categories. This is particularly important for the end-point categories, as the respondent should know what an extreme score represents.

Midpoints on summated-rating scales also seem to cause a great deal of concern for some scale developers. A scale with a defined midpoint is one such as the following.

Strongly Disagree	*Disagree*	*Undecided*	*Agree*	*Strongly Agree*
1	2	3	4	5

A scale without a defined midpoint would read like the following.

Strongly Disagree	*Disagree*	*Agree*	*Strongly Agree*
1	2	3	4

There are no strong arguments from a statistical perspective for one type of response scale over the other. The issue that I have found to be most relevant is how the respondents will deal with no midpoint. With no midpoint, the effect is to force respondents to choose to agree or disagree with a statement about which they may be truly ambivalent. I have had the experience where respondents became frustrated with no midpoint and refused to complete the scale or refused to answer the item. Sound advice on this issue is to make a rational decision as best you can. If most people completing the scale will have feelings of agreement or disagreement about virtually all of the items, then no midpoint will stop those who "sit on the fence" from circling the midpoint all the time. If at least some of the respondents are expected to be undecided about some of the items, then it is better to provide a midpoint as it is a more accurate reflection of their attitudes, and they will be more likely to complete the scale.

Another of the questions about response scales is whether to include a "don't know" or "not enough information to make a judgment" type of response. Some would argue that these should not be included because it gives respondents an "out." Others argue that if respondents really do not know the answer, they should be allowed to indicate that on the scale. For example, suppose a group of workers is asked to rate their agreement with the following item: "My organization's goals are aligned with my own values." If the workers don't know the organization's goals, how will they know if the goals are aligned with their values? In this case, a "don't know" response is very useful. In item construction, try not to include items to which individuals will not know the response. When it is anticipated there may be some "don't knows," adding the option is a wise move.

Another issue that comes up more in research in the field than in lab-based research is use of items with a negative valence. These are the items that need to be reverse coded before summing across the items. For example, items with a positive valence about job satisfaction might include "my job is an exciting one," "my supervisor listens to my suggestions," and "my coworkers are supportive." Then the scale developers throw in an item such as, "I have no commitment to my organization." If all of these were to be responded to on a five-point Likert-type scale with responses ranging from strongly disagree to strongly agree, high scores on the first three items would mean higher levels of satisfaction. High scores on the fourth item would indicate low levels of satisfaction.

In scale construction, it is frequently advised to include items that are negative in valence to ensure that the respondent is paying attention to the items. It prevents respondents from always selecting a particular response category without really attending to the item. While these are logical reasons for including negative valence items, they are also problematic. Specifically, I have found two issues arise. The first

is that respondents who are not students in university classes don't like these types of items. It has been reported back to me that the negatively worded items are confusing. In addition, on more than one occasion, I have had to delete the items because analyses of responding patterns suggested that the respondents made mistakes on these items. I am not alone in these observations (e.g., DeVellis, 2003; Netemeyer, Bearden, & Sharma, 2003).

Therefore, it is advisable to use negatively worded items with caution. Be deliberate in making the decision of including such items. If the sample of respondents is under time pressure to complete the surveys, not used to completing surveys regularly, or might be easily confused, then don't use them. If it is decided to use such items, include many of them. That is, don't write 19 positively worded items and then add one that is negatively worded. Instead, write 10 positively worded and 10 negatively worded items so that the respondents become used to the fact that negatively worded items are a usual occurrence on the survey. Finally, make sure that the negatively worded items are interspersed throughout the survey and not all at the beginning or end.

Response categories should also allow for an even distribution of negative and positive attitudes or else they will negatively or positively skew the interpretation. Here is a typical example: A marketing representative calls me and asks me about the customer service I received when I took my car in for service. I'm asked to rate the service on the following scale: "unsatisfied, satisfied, very satisfied, or extremely satisfied." The problem here is that only one of the four options allows me to express a negative attitude; the other three are levels of satisfaction. Be suspicious of companies that say they have a 90% customer satisfaction rating. To interpret this statistic, it would be necessary to see the questions *and* the response options.

Other Types of Continuous Response Scales. Visual analogue scales are a variant on the multicategorical approach to responses. In these scales, the respondent is presented with an item and asked to make a mark on a line, and the score for that item is the number of millimeters from one end point. Although there are no inherent statistical problems with this approach, it is quite time consuming to score. Visual analogue scales are common in assessments of health and stress. Figure 3.2 shows a response of this nature.

If a respondent was asked to mark an X on this scale to show the level of stress he or she was currently feeling and marked it as shown, the score on this item would be the distance from the left end point to the X. Higher scores would correspond to higher levels of stress.

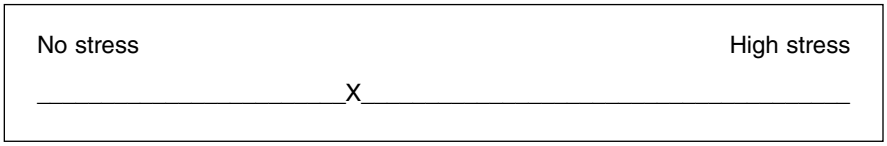

Figure 3.2 Visual Analogue Scale

Figure 3.3 Facial Scale

Another response format is pictorial. These are particularly useful for respondents who do not have strong verbal skills, such as young children, people without language proficiency in the language in which the test was constructed, those with low literacy, and so forth. The items can be read out loud by the test administrator and test takers respond to the items by selecting the facial expression that best captures their attitudes. Figure 3.3 shows an example of a three-alternative facial response format. A question might be posed to the individual: "How do you feel about your psychometrics class?" The individual then would be expected to choose the facial expression that most closely matches his or her affective reaction to the psychometrics class.

Adjective rating scales are another way to obtain continuous responses. In these scales, the ends of the scale are anchored with presumed polar opposites. These are called *polar adjective* rating scales. An example of one follows:

1. quiet	____	____	____	____	____	loud
2. sincere	____	____	____	____	____	insincere
3. happy	____	____	____	____	____	sad
4. selfish	____	____	____	____	____	selfless
5. ambitious	____	____	____	____	____	lazy

An X that is placed on the line in between each polar pair is the respondent's scale value for that item. If a respondent placed an X on the third line for Item 1, it would indicate that the individual thought he was neither quiet nor loud but right in the middle.

Some of the scales that have been created this way make assumptions about the oppositeness of the adjectives. A good example is in the above list. While most people would agree that loud and quiet are opposites, as are happy and sad, it is not as clear that ambitious and lazy are necessarily opposites. Thus, when using or creating scales like these, it is critical that the decisions made to create the opposing pairs are defensible.

Another important contribution to the continuous scaling process has been in the design of category descriptors. These are often referred to as *anchors.* As with

the Behavioral Observation Scale (BOS) scaling process discussed in Chapter 2, the behaviorally anchored rating scales (BARS) process also uses a critical incident technique and can be used to create the anchors for such scales. One of the most frequent uses of BARS is in the assessment of employee performance, but it can easily be used in a number of other settings.

The process for developing anchors for BARS in a work performance setting would be to first have subject matter experts (SMEs) generate a list of critical behaviors that result in good and poor performance. For example, good performance for a sales staff employee might include approaching customers coming into the store within 30 seconds, smiling and saying hello to customers, asking customers if there is anything in particular that they are looking for, and so forth. Poor performance for a sales staff employee might include waiting for customers to approach the staff member to secure assistance, not smiling at customers, not offering to assist customers to find what they need, and so forth. All of these behaviors are associated with customer service. Other domains of behaviors would also be generated for such things as assisting other sales staff, punctuality, flexibility in scheduling, solving customer service problems, and so forth.

Next, the behaviors are clustered together into their domain content areas. For example, all the good and poor behaviors generated that were associated with customer service would be grouped together. All the good and poor behaviors that were associated with helping other sales staff would be grouped together, and so on. To check whether the behaviors were correctly grouped into their domain content area, all of the behaviors for all of the domains would be gathered and mixed up, and another sample of SMEs would be asked to sort them into their domain areas (e.g., customer service, assisting other sales staff, etc.). These new SMEs would be expected to show consistency in grouping the behaviors into the same domain areas as the first set of experts.

An agreement percentage is set in advance to determine if a particular behavior is confusing. For example, if a behavior such as "approach customers coming into the store within 30 seconds" is re-sorted by 90% of the new SMEs into its original domain of customer service, then that behavior would be said to be consistent. On the other hand, if a behavior such as "counts out correct change for the customer" is re-sorted by 50% of the new experts into its original domain (e.g., completes sale accurately) and 50% into another domain (e.g., customer service), then that behavior would be said to be inconsistent, and would likely not be used in the final scale. The percentage agreement level the scale developer sets is really up to him or her. Previous research in the same area, as well as common sense, is useful as a guide in selecting a percentage agreement rate that is appropriate.

Next, each SME is asked to rate the effectiveness of the behaviors that survived the sorting process in terms of the job performance domain (usually on a five-, seven-, or nine-point scale). Behaviors that have high variability in the ratings are dropped. For example, if a behavior such as "approaches customers within 30 seconds of entering the store" is rated as a 6 or a 7 (i.e., highly effective) on a seven-point scale by all of the SMEs, then the behavior is consistent. If half of the experts rate the behavior as a 3 and the other half as a 7, then the variability of the ratings is quite high. This item, then, would likely not be retained. The specific

level of variation selected as the cutoff is up to the scale developer, but should be consistent with previously developed scales in the area. Finally, the scale is developed so that the domain of "customer service" can be assessed on a three-, five-, or seven-point Likert-type scale with behavioral descriptors anchoring various parts of the scale. It is not necessary to anchor all of the points on the scale; however, the end points and the center point of the rating scale must be clearly anchored.

Intensity Versus Frequency Likert-Type Scales. One general issue that does arise in creating Likert or Likert-type scales is whether to ask the respondent for intensity information, such as levels of agreement, liking, or satisfaction, or whether to ask the respondent for frequency information, such as, "How often or how frequently do you experience headaches?" or "How often do you observe your coworker coming in to work late?"

The following is an example of an item asked in an intensity manner. "Rate your level of agreement (on a 1–5 scale) with the following statement: I am happy most of the time." The item could also be asked in a frequency manner such as the following: "How often do you feel happy?" The respondent has to read the category descriptors carefully to see which one is most appropriate. Assume the response options are "never," "sometimes," "frequently," and "almost always." The problem is that what the test developer meant by *sometimes* may very well not be what the respondent means by *sometimes.* In fact, each and every respondent may have a different interpretation of *sometimes.*

Clearly, there are times when frequency is an important measure. For example, rating the frequencies of behaviors or symptoms is often of great import in both diagnosing a disorder/disease and assessing the effectiveness of treatments. This would be the case for physical and psychological symptoms, behavior problems, and so forth. Frequencies are also of import in less life-affecting areas, but ones where frequency measurement is appropriate—for example, assessing an aspect of employee performance by counting the number of errors made, or assessing safe driving behavior by counting the frequency of carrying out safe actions.

However, it is critical that the category descriptors for such scales are carefully constructed. An appropriate set of category descriptors for the question "How often do you feel happy?" might be the following: "less than once a week, once a week, two to three times a week, four to five times a week, or more than five times a week." Test constructors must know the construct under investigation quite well in order to create the correct category descriptions. That is, for any given item, they need to know the typical frequency, very low and very high frequencies, and moderately low and moderately high frequencies. These, then, become the anchors for the points on the scale.

Ipsative Versus Normative Scales

Up to this point, the focus of scale development has been on normative scales. In normative scales, there is one measurement scale for every construct of interest—whether that be an attitude, personality measure, interest, or ability. The usual normative scaling development and use procedure is to create a scale, ask a number of

different individuals to complete the scale, compute each respective score, and then use the scores for direct interpretation (e.g., comparing one to another, one to a group, one group versus another) or in relating the scores to other variables of interest. This is also referred to as *nomothetic assessment.*

It is assumed in normative scales that the scores underlying the construct being assessed are normally distributed in the population. Thus, normative scores are used with normative scales to compare *across* individuals. In ipsative scales, there is a separate scale for each individual respondent, and the population of that individual's trait scores is distributed about the mean of that individual's scores. Thus, the purpose in ipsative scoring is to make comparisons about different constructs *within* each individual. This is also referred to as *idiographic assessment.*

The following are some examples of ipsative items:

1. Rank order the following ice cream flavors (vanilla, chocolate, strawberry, butterscotch), with 1 being the most tasty and 4 being the least tasty. (If the first three flavors of ice cream are ranked, then it is known by default what the fourth ranking is going to be.)
2. Which of the two items best describes your interests: "I like to arrange flowers" or "I like to solve computer software problems"? (If flower arranging is selected as the interest, then it is known that the respondent is less interested in solving computer software problems.)
3. Rank order the four statements as to their description of your supervisor, with 1 being most like and 4 being least like your supervisor:

   "My supervisor always asks for my opinion on matters that affect my work,"

   "My supervisor frequently asks for my opinion on matters that affect my work,"

   "My supervisor sometimes asks for my opinion on matters that affect my work," and

   "My supervisor rarely asks for my opinion on matters that affect my work."

   (If rankings for three of the supervisor descriptions are provided, then it is known what the fourth one is going to be.)

All of these items have a similar characteristic—knowledge of the response to one item provides information about what the respondent will (and can) put for another item. This means that the responses are not independent of one another. The lack of item independence makes ipsative data inappropriate for many analyses that make an implicit (and sometimes explicit) assumption about item independence.

Respondents often report a dislike of ipsative measures. This is because they may not be able to accurately make the judgments requested. For example, what if respondents like vanilla and chocolate ice cream to exactly the same degree? By forcing them to rank order the flavors, an accurate assessment of ice cream flavor liking is not possible. What if respondents don't like flower arranging or solving computer software problems? This dislike cannot be captured. What if a respondent's supervisor rarely asks for opinions so this option is ranked 1, but all the rest are equally unlikely? If that is the case, the 1–4 ranking will not capture the reality of the supervisory situation.

From a statistical perspective, many researchers have gone so far as to suggest that purely ipsative measures have such severe limitations that they should not be used (e.g., Cornwell & Dunlap, 1994; Hicks, 1970; Tenopyr, 1988). A special factor-analytic approach called the *Q-technique* is needed when scores are ipsative (Guilford, 1952).

There are several reasons for not using ipsative measures; however, there is evidence that ipsative scores are useful in some respects (e.g., Greer & Dunlap, 1997; Ravlin & Maglino, 1987; Saville & Willson, 1991). The primary use is when intraindividual assessment is the goal of the testing situation. For example, if an individual desires information about his or her most appropriate career options, forcing the respondent to choose between activities associated with vocational alternatives provides valuable information for that one individual. That is the basis for the Kuder Occupational Interest Survey (Kuder, 1979). Kolb (1985) has been assessing preferred learning styles with his ipsative measure for many years. Assessment of various aspects of personality with ipsative measures has also been conducted—with the Edwards Personal Preference Inventory (Edwards, 1959) and with the Myers-Briggs Type Indicator (Myers, McCaulley, Quenck, & Hammer, 1998).

One variant on the purely ipsative versus purely normative scaling is a particular type of forced-choice item and summing process. In purely ipsative scales, the response choices are pitted against one another so that choosing one option by definition makes the respondent higher on one scale and lower on the other. So, for example, if I choose an option that puts me higher on the extroversion scale, I am by default lower on the introversion scale. Some forced-choice formats are created so that choosing one option does not by default ensure a lower score on the other. That is, respondents are asked to respond using a forced-choice format, but the item responses are independent of one another. An example of this type of scale is the Sensation Seeking Scale (Zuckerman, 1979). It is a 40-item scale that has four subscales, each of which is made up of 10 forced-choice questions. The individual respondent can be high on all four subscales, low on them all, or any combination thereof. These types of forced-choice scales seem not to suffer the statistical problems of the purely ipsative scales.

Another type of approach to generating both normative and ipsative data is with the Q-sort method. This has been used in personality assessment (Block, 1978) and in assessing organizational value congruence (Chatman, 1989). A series of items (say 50) is generated to capture the domain of interest. Next, an individual sorts the items into a specified number of categories (say 10) on the basis of the importance or relevance of the item for the domain. For example, assume that the domain of interest is organizational effectiveness and the purpose of the task is to compare a potential job applicant's values to those of the organization's. If an organizational attribute such as "allows for participation in decision making" is very important to the applicant in the domain of organizational effectiveness, he or she might place that item in the 9 or 10 category. If "competitive pay" is of moderate importance to the respondent in the domain of organizational effectiveness, the item might be placed in the 5 or 6 category. The applicant does this for each of the 50 items. Fewer items are permitted in the extreme categories (e.g., 1, 2, 9, and 10). The result is supposed to be a normal distribution of items such that most are piled up in the more

neutral categories and then taper off to the ends. This produces an ipsative set of values for the applicant.

The next step is to have a large number of incumbent employees do the same sorting but with the caveat that they sort on the basis of how important each item is to the organization as it currently operates. These current employees thus provide a distribution of organizational attributes that range from more to less important. These employee-sorted item category assignments are averaged and this produces an average of ipsative data that represents the importance of various organizational attributes.

Now the applicant's profile can be compared with the organizational profile by correlating the item category assignments the applicant made with the category assignments made by the large employee group. The higher the correlation, the more likely the person will "fit" the organization. This correlation provides a normative set of values that can then be compared across individuals.

So, a question remains about what type of scaling is best, ipsative or normative? As with many of the choices that have been posed thus far, there is a need for making a reasonable decision based not on statistical grounds alone, but on rational grounds. The most important issue is what the information will be used for. If a set of items has been carefully developed and shows strong psychometric characteristics in normative samples, revising the items into an ipsative measure can force high degrees of intraindividual variance, which is often useful in describing one person. This type of information is most helpful when the individual appears for assistance (for example, in vocational counseling). If the purpose is to compare across individuals, then clearly a normative scaling scale is most appropriate.

Difference and Change Scores

Another debate that has raged in the psychometric literature is over difference scores. Often it is of interest to researchers and practitioners alike to ask questions such as "How much autonomy do you have in your job?" and "How much autonomy would you like to have in your job?" The respondent rates his or her degree of autonomy for the first at a 3 on a seven-point Likert scale. Then he or she answers 6 to the second item. The difference is 3 points on the scale and would indicate that the person would like more autonomy than is presently experienced. The difference is between two conceptually linked but distinct constructs (called *components* in this literature).

Change scores assess the same individual using the same measure but at two (or more) different times. Researchers and practitioners in the areas of education and evaluation would be likely to use change scores as data points. They are not the same as difference scores, but both change and difference scores have come under fire. The issues with difference scores will be discussed first.

There are problems with simple difference scores, and the three most common will be discussed here. The first concern is that the reliability of difference scores will most likely be lower than each of the component scales that make up the difference score. Another problem is that they do not account for any variance in a criterion above and beyond that accounted for by each of the components. Assume everyone in an organization ($N = 2{,}000$) indicated on a five-point scale how much they felt the

organization valued their contributions (variable *X*). They are also asked how much they would like their organization to value their contributions on a 1–5 scale (variable *Y*). Next, the difference between these two values for all 2,000 employees is calculated (variable *Z*). A criterion variable such as job performance is then collected on all of the employees as well. Opponents of using the difference score argue that if the job performance measure is regressed on *X* and *Y, Z* will not add any additional information above and beyond that provided by *X* and *Y* separately.

The third criticism of difference scores is a conceptual concern rather than a statistical one—just what is it that the difference measures? Researchers and practitioners using difference scores must pay close attention to what they are measuring. For instance, difference scores have been calculated in various ways: algebraic differences (this means that the signs of the differences are left intact), absolute differences, and squared differences are the most common. In many instances, the rationale for using one over the other is not made. Sometimes it is critical that the sign, or direction of the difference, is part of the construct. At other times, the magnitude of the difference, regardless of direction, is important.

If difference scores are used, addressing the psychometric concerns that have been expressed is necessary. The component scales must have good reliability and there should be high variability on each of them. In addition, the components should *not* be highly correlated with one another. In the unusual instance when the component scores are negatively correlated, the reliability of the difference score will actually be higher than the reliability of the two component scales. Component scales that use multiple items and/or are completed by two different sources (e.g., employees and managers) are more likely to have more reliable difference scores than component scales that use single items and/or are both completed by the same person. Be thoughtful about how difference scores are constructed and what they will mean. Make sure that they add incremental information above that contained in the component scores. Be aware of the arguments for and against using difference scores by reading some articles on the subject matter and make a case for why they are being used. Some excellent articles on this issue include Edwards (1993), Edwards (1994), Edwards (1995), Edwards and Cooper (1990), Johns (1981), and Tisak and Smith, (1994).

Change scores, as noted earlier, are the same measure used on the same individuals taken at different times. For example, suppose a sample of students is measured at the beginning of a typing course on the number of words typed correctly per minute. The students are then sent through 6 weeks of typing training. After the course, they are measured on the number of words typed correctly per minute. The difference between the two measures is a change score, or gain score, as it is sometimes called.

The reliability problem alluded to in the difference score literature is the same in the change score literature. In fact, this is the crux of the argument about why not to use change scores. Much research and argument abounds in the psychometric literature about change scores. References that are highly useful in understanding the arguments include Collins (1996); Cronbach and Furby (1970); Humphreys (1996); Rogosa and Willett (1983); Williams and Zimmerman (1996a); Williams and Zimmerman (1996b); and Zumbo (1999). Other readings are more helpful in that they provide alternative approaches to the simple change score, and these include Collins and Sayer (2001); Cribbie and Jamieson (2000); Rogosa, Brandt,

and Zimowski (1982); and Tisak and Tisak (1996). One particularly helpful approach has been clearly described by Zuckerman, Gagne, Nafshi, Knee, and Kieffer (2002). They make a compelling argument that one can defensibly create a difference score using measures of the same construct (e.g., two different measures of need for achievement). However, if the difference score is generated from two different constructs (e.g., the difference between actual and ideal organizational attributes), then an interaction term should be created from the two component measures first. Then, any relationship of the interaction with a criterion should first take into account the two component measures.

Change scores on a typing test are based on psychomotor skills and are problematic from a reliability standpoint. Change scores that use attitude measures are problematic for another reason; they sensitize the test taker to the issue at hand. For example, suppose a measure of employee job satisfaction is taken at one time, then an intervention is introduced (e.g., managers are trained to be more sensitive to employee needs), and then employee job satisfaction is measured at a later time. The employee scores are likely to change from the first to the second administration of the job satisfaction measure. However, employees have been sensitized to the issue simply by having completed the first job satisfaction measure. Employees may have had a heightened sensitivity to any changes in the job environment and these may have artificially inflated their job satisfaction scores at a later time. This issue of sensitization needs to be taken into account when interpreting change scores.

It is worthwhile to be familiar with the change and difference score literature when using a change score research design. Williams and Zimmerman (1996a) note that there are assumptions in the attack on change score reliability that may not be met. Specifically, the attack assumes the worst-case scenario, where the variances of the pretest and posttest are equivalent and where the correlation between the pretest and posttest are high. The reliability of a change score increases as (a) the correlation between the pretest and posttest decreases and (b) the ratio of the variances of the pretest and posttest deviate from 1.0. If the data show that the pretest and posttests are highly correlated and they have similar variances, consider an alternative to using the change scores.

Summary and Next Step

In this chapter, the many issues facing scale designers in structuring the type of response desired were covered. These included

a. deciding whether to use an open- or closed-ended format;

b. when choosing a closed-ended format where there is a "correct" answer, settling on the number of alternatives, carefully crafting distractors, determining if there will be a penalty for guessing, choosing whether or not to use a power or a speeded test, and deciding what will be done with omitted responses and partial credit;

c. introducing the revolutionary work of Rensis Likert in attitude assessment;

d. in attitude assessment, discussing the purpose of the instrument, decisions about the type of scale to use (frequency versus intensity, ipsative versus normative), the number of alternatives to use, and the anchors/descriptors for the categories; and

e. reviewing issues that have arisen based on calculating difference and change scores and the fact that the pitfalls associated with these are worth attending to before trying to publish the scale and work associated with the scale.

Most of the literature on scale development focuses on the issues of creating samples of items from the population of items and, to this point, this has been the focus. However, sampling issues as they pertain to respondent populations are just as important, so in the next chapter issues of respondent samples are covered.

Problems and Exercises

1. Write an open-ended question asking your colleagues for information about their work experience. Ask a few of them to respond to the items.
2. Write some closed-ended questions about your colleagues' work experience. Ask a few of them to respond to the items.
3. Assume you ask 50 students to indicate where they have had the majority of their work experience. You get the following numbers: retail = 25, food service = 15, financial sector = 7, and other = 3. Calculate if there is a proportional difference in the data.
4. Assume you have the same data as in Problem 3, but now you break it down by age (those 25 years and under and those 26 years and older). You obtain the following table of data (Table 3.7). Calculate the chi-square statistic and Cramer's V for this table. Interpret your results.

Table 3.7 Cross-Tabulation of Age and Employment Frequencies, Row and Column Totals, and Row Percentages

	25 and Under	*26 and Older*	*Row Total*	*Row %*
Retail	16	8	24	48%
Food Service	10	2	12	24%
Financial	3	8	11	22%
Other	1	2	3	6%
Column Total	30	20		

5. Indicate if the following are true or false dichotomies (as measured in brackets):
   a. Gender (male/female)
   b. Age (under 30 years/over 30 years)

c. Marital status (married/not married)
d. Job attitude (satisfied/not satisfied)
e. Organization (organized/disorganized)
f. Student status (registered student/nonregistered student)
g. Speed (over the speed limit/under the speed limit)
h. Incarceration (in jail/not in jail)

6. Take a subject material (stimuli) such as a pizza. Generate an adjective checklist associated with that stimulus. Ask four of your colleagues to check off the items they think characterize the stimulus.

7. Using the material in the first three chapters of this book, write a multiple-choice test item with three alternatives (one correct and the other two incorrect). Ask your classmates to complete the test item. Generate a table of responses to the item. Using the test of single proportions, calculate whether the correct answer or one of the "foils" was more likely to be selected.

8. What would be the "corrected for guessing score" for someone who answered 45 questions out of 50 correct on a true/false test? What would be the "corrected for guessing score" on a multiple-choice test with four alternatives to each item?

9. What are the differences between power and speeded tests?

10. Assume I take a 100-item multiple-choice test that has three alternatives for each item. I get 80 correct, but skip five of the items. If my instructor gives omitted credit for the five items, what would be my score?

11. Using the construct you have been developing thus far, create five Likert-type items labeling your response categories.

12. What is an item with a negative valence? What are the arguments for including and arguments for not including such items?

13. What is a visual analogue scale and how is it scored?

14. Why are pictures used sometimes as responses?

15. Create a polar-adjective scale for your stimuli from Problem 6. Try to generate at least five pairs of polar opposites.

16. What is the utility of a behaviorally anchored rating scale?

17. Using the construct "healthy lifestyle," generate five items that ask for frequency responses on a Likert-type scale.

18. What is the difference between normative and ipsative test items? Why would one use one type versus the other?

19. What is a difference score and what are the arguments that have been leveled against using such scores?