CHAPTER 2

Designing and Writing Items

Once the construct (or constructs) to be assessed is clarified, the next step is to get on with the business of writing items that assess various aspects of the construct. The approaches to designing items have been compartmentalized into three general categories: empirical, theoretical, and rational. However, this compartmentalization is somewhat artificial, as the three approaches often overlap.

Empirical, Theoretical, and Rational Approaches to Item Construction

Empirically generated items are those that "do the job" but the test developer either does not know and/or does not care why the item is relevant. As an example, items for the Minnesota Multiphasic Personality Inventory (MMPI), developed and published by Hathaway and McKinley in 1943, were designed and retained because they were able to distinguish between two groups of individuals. These groups could be "normal" versus "paranoid" or "depressed" versus "schizophrenic" and so forth. Because the purpose of the scale was to distinguish between normal and psychiatrically disturbed individuals, the criterion for retaining an item was the degree to which it was able to make the distinctions required, regardless of the item's content. The MMPI comprises a total of 550 statements, to which the respondent indicates "yes," "no," or "cannot say." There were nine clinical scales and four validity scales in the original MMPI. Other scales have been added since it was created. This "dust bowl" empirical approach characterized many scales that were developed in the middle of the 20th century.

A different tactic used to design test items is a theory-driven approach. An example of such a scale is the Wechsler Adult Intelligence Scale-III (WAIS-III;

Tulsky, Zhu, & Ledbetter, 1997). In 1939, David Wechsler designed and published a test of intelligence aimed at the adult population. There have been several revisions to the test, the most recent in 1997. His theory of intelligence was that it had different facets, each of which should be assessed with different methods. For example, the Information subtest is made up of general information questions that can be answered in a few words. The Picture Completion subscale shows pictures on cards, each with a missing part, and the test taker must identify the missing part. The Digit Span subtest asks the test taker to repeat strings of digits forward and backward. Other subtests on the WAIS-III include Picture Arrangement, Vocabulary, Block Design, Arithmetic, Object Assembly, Comprehension, Digit Symbol, Similarities, Letter-Number Sequencing, Matrix Reasoning, and Symbol Search. You can see that Wechsler had a theory about intelligence (that it was multifaceted), created items to assess each aspect of that facet of intelligence, and asked the test taker to make responses that he believed were relevant to each facet of intelligence.

Rationally developed scales use a combination of theory and empirical work to guide the item construction and retention. A good example of a rational scale is the Jackson Vocational Interest Survey (JVIS; Jackson, 1977). Jackson initially created hundreds of items for his survey. The items were based on the vocational literature, and he used various statistical procedures to whittle the items down to the ones that he found to be most useful. In the end, he generated a scale that is praised as being one of the most carefully constructed scales available. The JVIS measures 34 basic interest scales, 26 work role scales, and 8 work style scales. Most scales that are developed now use a rational model; they are not likely to be purely empirically based nor purely theoretically based.

Literature Search. The science side of item development provides a pretty clear roadmap of what is expected in this phase. The first thing to do when creating a scale is to go to the theoretical and empirical literature that will have an impact on the scale's development. Using the team player construct described in Chapter 1, for example, how other researchers have defined (both conceptually as well as operationally) being a team player as well as potentially related constructs has to be examined. In this instance, journals that deal with workplace issues—such as *Journal of Applied Psychology, Journal of Occupational and Organizational Behavior, Academy of Management Journal,* and *Administrative Science Quarterly*—will be reviewed to see which articles include the words *teams* or *team player.* Journals that contain articles about groups and personality in broader settings, such as *Journal of Personality and Social Psychology, Journal of Small Group Research,* and *Journal for Specialists in Group Work,* are also potential sources of information. Journals that specialize in scale development and assessment, such as *Assessment, Educational and Psychological Measurement, Psychological Assessment, International Journal of Testing, Journal of Psychoeducational Assessment,* and *Journal of Educational Measurement,* may also provide ideas about how others have attempted to measure the construct.

Books on teams in organizations abound. Some have an empirical approach, some are more theoretical in orientation, and some have an anecdotal flavor.

Regardless of their approach, they need to be examined for how the authors have defined the construct of being a team player.

The World Wide Web provides links to several test locator services. Searching these sites for existing instruments is a very fast and potentially fruitful activity. The following websites can be helpful: www.unl.edu/buros and www.ets.org

If there is existing research where the authors already have gone to the trouble of developing items for such a construct, consider using that research rather than developing new items. This will depend on how close a match there is between the construct that you want to assess and what the developed scale assesses, how carefully the items were developed for those instruments, and the results of the analyses that assess the instrument's psychometric properties. It is always easier to use an existing scale or even modify an existing one than it is to create a new scale.

If you do decide to modify a scale by deleting items, rewording items, changing the response format, or making any other change, the psychometric properties of the new scale based on analyses with the new sample will need to be reported.

Subject Matter Experts. The next place to go for information about your construct is to subject matter experts (SMEs). SMEs come from a variety of settings. One obvious group is made up of those individuals who study the phenomenon that you are interested in measuring. These individuals will be easily located after having completed the literature search. Names of some researchers will keep recurring, and these are the people you will want to contact. They are usually found in university and college settings.

They are often very helpful and willing to assist in the construction of your scales, particularly if you've done your homework and found out a lot about the subject matter through your literature search. You may want to interview these SMEs in person, by phone, or electronically to ask them about the construct you want to measure. They will often refer you to new sources that you may not have come across. They may refer you to other researchers and/or graduate students working in the same area, thus providing you with access to more and more individuals who know about the construct.

Another group of SMEs would be laypersons who have specific knowledge about the construct you are interested in assessing. For example, if we want to find out information about what is important to being a team player in a work setting, we would likely want to speak with individuals who work on or with teams. Interviews with people who work on teams, manage teams, and facilitate teams in work environments will provide a perspective that the researchers cannot on what it means to be a team player. Again, interviews with these SMEs are invaluable to understand as clearly as possible what it is you want to measure.

A familiar question in terms of interviewing SMEs is, How many should I interview? The answer is, As many as it takes until no new perspective or information is obtained. Those who use qualitative data collection procedures on a regular basis call this point the *saturation of themes.* Let's suppose I interview some top researchers in the "team" literature and I find that in my fifth interview, I no longer get any new information on what it means to be a team player. I decide to carry out another two interviews just to make sure that I don't miss anything important.

If the sixth and seventh interviews also yield no new information, it is likely that my themes are saturated.

Then I would turn to my practitioner SMEs and interview people who work on different types of teams in different organizations, people who are managers of work teams, and people who are called in to work with teams when they are experiencing difficulty. If I use the same criterion of theme saturation, I may find, at the end of my theme saturation approach, that I have interviewed 10 people who work on teams, 9 team managers, and 8 team facilitators. Box 2.1 shows an example of how theme saturation would work.

In my own experience in developing scales used in research (e.g., Kline, 1994; Kline & McGrath, 1998; Kline, 1999; Kline, 2001; Kline, 2003) as well as in practice (e.g., Kline & Brown, 1994; Kline & Brown, 1995; Rogers, Finley, & Kline, 2001), I have found that after about seven interviews, I don't usually get any new information. The important consideration here, though, is that the seven interviews are with only one constituent group. So, for example, if I want to develop an instrument to assess teacher effectiveness in a psychological testing course, several constituent groups would have perspectives to contribute: students, instructors, teaching facilitators, and administrators. It is likely that I'll need to conduct 7–10 interviews with each constituent group.

Sometimes the constituent groups don't add any real new information, but need to be included for political reasons. For example, I was involved in developing a scale to assess the stress of city bus drivers (Angus Reid Group, 1991). The project team interviewed bus drivers of varying seniority, dispatchers, managers, bus riders, and members of the city council. For the report to be believable and the recommendations acted on, the constituent groups wanted to see that their perspectives were heard and incorporated into the survey.

Once the themes for a construct have been identified, items need to be created that transform the construct into something that can be measured with a number attached to the level of the construct. The next part of this chapter is devoted to how to go about that task.

Writing Items: Guiding Rules. Continuing with the team player example, we should now have a wealth of information about what it means to be a team player. This information has been obtained from a variety of sources and thus we can be fairly confident that, in Western-culture organizational settings, we know what being a team player means and have a sense of how to operationalize this construct through measurement.

So how are items written? Depending on whether the scale is to assess an attitude, content knowledge, ability, or personality trait, the types of issues will be somewhat different. One consideration is to determine whether the items should be questions or declarative statements. Regardless of what type of item is chosen, stick with the pattern for the entire instrument. For example, do not shift from declarative to interrogative statements.

The overall issue in item writing is one of clarity—be as clear as possible in asking respondents the questions. The more clear the question, the more confident

Box 2.1 An Example of Theme Saturation Process for an Aspect of Team Player Predisposition

One question that would likely be asked of SMEs about their experiences working on teams is, Give me an example of a time when a team you were on was a positive experience. Here are some examples of some responses you might get to such a question.

SME 1: I was a member of a soccer team as a kid. Whether we won or lost, the coach always had something good to say about each and every player and pointed out something for each of us to work on for the next time. This made us all feel like she cared about us as individual players.

SME 2: When I was in high school, I was in a school play. We rehearsed for weeks and everyone appreciated everyone else's accomplishments and input into the play. We all knew that, down to the ticket takers, the play would not be a success unless we all pulled together. On opening night, it was a really wonderful feeling to hear the applause at the end of the show.

SME 3: I worked at a retail clothing store once where the shift on the floor was considered a team. Sales commissions were not given out for each person but for the whole shift on the floor at the time. This worked really well for me because I was not very outgoing and was shy of approaching customers. However, I was always there to help the salesperson assisting the customers to get different sizes and to check them out at the end. This freed up the sales staff that were more outgoing to play to their strengths and approach new customers.

SME 4: I worked on a group project in a college English course where the assignment was to review how a number of different writers described their experiences immigrating to a new country. We each quickly decided as a group how to analyze the content and style of the writer. Then we decided who would research what writer and how we would integrate the overall conclusions. Everyone pulled their own weight and we learned a lot from each other.

SME 5: My experience as a team member was when I was a flight crew member for an airline. There were five of us including the pilot, copilot, and navigator. All of us would greet the passengers as they boarded the aircraft. The pilots would know all of our names and would treat us with a great deal of respect, even though they had a much higher rank in the organization. Our purpose was clear—we were there to ensure the safety, comfort, and enjoyment of the passengers.

Although usually you would continue to interview others, it is clear already that some of the themes that recur in these answers are the following:

1. Each team member contributes.
2. Each team member recognizes the contributions of other members.
3. Team leaders treat all members of the team as valued contributors and with fairness and respect.
4. The purpose of the team is clear.

These themes can then be converted to items such as the following:

My experiences about being a member of a team include

1. each team member putting forth equal effort,
2. each team member knowing the teams' purpose,
3. team leadership that was respectful,
4. team leadership that was fair, and
5. team members appreciating the efforts of all members.

you'll be that the respondents have provided the information desired. There are some guiding principles that are quite obvious but often not attended to in writing items. These principles are based on the guides cited in Ghiselli, Campbell, and Zedek (1981) and Nunnally and Bernstein (1994).

1. Deal with only *one* central thought in each item. Those with more than one are called *double-barreled.*

   Poor item: My instructor grades fairly and quickly.

   Better item: My instructor grades fairly.

2. Be precise.

   Poor item: I receive good customer service from XYZ company.

   Better item: A member of the sales staff at XYZ company asked me if he or she could assist me within one minute of entering the store.

3. Be brief.

   Poor item: You go to the corner store and decide you want to buy $10 worth of candy. You see that you have only a $20 bill and so pay with that. How much change would you get back?

   Better item: If you purchase $10 worth of candy and pay with a $20 bill, what change would you receive?

4. Avoid awkward wording or dangling constructs.

   Poor item: Being clear is the overall guiding principle in writing items.

   Better item: The overall guiding principle in writing items is to be clear.

5. Avoid irrelevant information.

   Poor item: Subtractive and additive color mixing processes produce differing results. If blue and yellow are added together using subtractive color mixing, what would be the resulting color?

   Better item: In subtractive color mixing, blue and yellow added together make . . .

6. Present items in positive language.

   Poor item: Which of the following does *not* describe a characteristic of a democracy?

   Better item: Which of the following is a characteristic of a democracy?

7. Avoid double negatives.

   Poor item: For what age group will it not be the case that people disapprove of graded vehicle licensing?

   Better item: What age group is most in favor of graded vehicle licensing?

8. Avoid terms like *all* and *none.*

   Poor item: Which of the following never occurs . . .

   Better item: Which of the following is extremely unlikely to occur . . .

9. Avoid indeterminate terms like *frequently* or *sometimes.*

   Poor item: Higher levels of self-efficacy frequently result after . . .

   Better item: Research shows that significantly higher levels of self-efficacy result after . . .

How Many Items? The question of how many items to include on a scale is simple to answer but difficult to pinpoint. The answer is, As many items as are necessary to properly assess the construct. The number will also be somewhat dependent on what types of analyses will be performed on the scale. It is unlikely that the construct of interest can be captured in a single item. Most analyses require at least 2 and more appropriately 5–10 items to perform analyses that suggest the construct is a reasonable one. Some analyses suggest that no fewer than 20 items for the construct is appropriate. However, 20 items to assess a construct may take a while for participants to complete. How many different constructs are to be assessed and how long it takes to respond to each item will play a role in determining how many items are appropriate for a given context.

Be cognizant of the mundane administrative issues such as how long the scale will take to complete. If it is too long, fewer people will be willing to respond. Those who do may be fatigued by the end and thus provide "garbage" answers. So the decision about how many items to construct and use has to be based on the rational consideration of both statistical needs and administrative concerns.

Attitudinal Items: Early Work in Item Generation

Burgeoning interest by social scientists in attitudinal assessment gave rise to various methods of developing and assessing numerical values for attitudinal items. Some of the assessments provided direct numerical estimates of stimuli differences and some provided indirect estimates. The indirect estimates need to be converted to numerical values.

Paired Comparisons. One of the first ways designed to assess item-to-response links entailed individuals making comparison judgments. Comparison judgments require the respondent to compare two different stimuli and make some sort of judgment about them. Paired comparisons require an individual to compare each stimulus with every other stimulus and make a judgment about their relative relationship. It provides an indirect assessment of the differences between stimuli (whether they be items or persons).

For example, I may have five graduate students that I am responsible for supervising, and I am asked to compare their performance. So, I have to compare Student 1 with Students 2–5, compare Student 2 with Student 1 and Students 3–5, and so forth. In total, I will have to make 10 comparisons. This highlights a drawback of paired comparison methods. That is, the number of comparisons goes up dramatically by adding more stimuli. If I have six students, I will have to make 15 comparisons. Be aware of this when using a paired comparison method as the

measurement instrument. The burden on participants becomes quite heavy as the number of stimuli becomes even moderate.

So what can be learned from data that are collected in a paired comparison method? To illustrate, here is an example of a paired comparison question: Is there a preference for a particular flavor out of four types of ice cream? Assume that 500 students are asked to make paired comparisons of the ice cream flavors: vanilla, chocolate, strawberry, and butterscotch. The data are then set up as in Table 2.1, where the proportions of people who preferred one type of ice cream over another are reported. First, note that the diagonal of the table is left blank, as this represents each ice cream flavor compared with itself. Second, note that the opposing proportions always sum to 1.0. That is because, if 0.25 of the students preferred strawberry to chocolate ice cream, then, by default, 0.75 preferred chocolate to strawberry.

The proportions are converted into a final preference scale by first changing the proportions to normal (z) values, summing down the columns, and then taking the averages of each sum. Table 2.2 shows the most preferred flavor is chocolate,

Table 2.1 Proportions of Paired Comparisons for Four Ice Cream Flavors

	Vanilla	*Chocolate*	*Strawberry*	*Butterscotch*
Vanilla		0.90	0.50	0.10
Chocolate	0.10		0.25	0.05
Strawberry	0.50	0.75		0.03
Butterscotch	0.90	0.95	0.97	

Note: The underlined values represent the derived proportions.

Table 2.2 Normal z Values for the Paired Comparisons for Four Ice Cream Flavors

	Vanilla	*Chocolate*	*Strawberry*	*Butterscotch*
Vanilla		1.28	0.00	–1.28
Chocolate	–1.28		–0.67	–1.65
Strawberry	0.00	0.67		–1.88
Butterscotch	1.28	1.65	1.88	
Total	0.00	3.60	1.21	–4.81
Average	0.00	0.90	0.30	–1.20

Note: The underlined values represent the derived proportions.

followed by strawberry, followed by vanilla, and with butterscotch a distant last. The resulting scale is an interval-level assessment of the degree of preference difference for ice cream flavor.

Items in Ranked Categories. One can also take stimuli and ask a number of judges to rank order the stimuli along some dimension. This is a direct estimate of stimuli although the scale is ordinal, not interval. The stimuli can be items in a test ranked on the dimension of "difficulty," they can be items in an attitude scale ranked on the dimension of "demonstrating altruistic behavior," or they can be students ranked on the dimension of "industriousness."

As an example, suppose I am asked to rank order the five graduate students I supervise in terms of which is most industrious and which is least industrious, with an award going to the most deserving. Further, suppose I ask three of my colleagues to do the same. Now I have a matrix of rankings like that shown in Tables 2.3 and 2.4.

Student 2 is rated top, followed by Students 1, 3, 4, and 5, in that order. So the award for industriousness goes to Student 2.

Table 2.3 Ranks of Five Graduate Students by Five Faculty Members on Industriousness

	Faculty Member (F) Rankings			
Student	*F1*	*F2*	*F3*	*F4*
1	1	2	2	2
2	2	1	1	1
3	3	3	4	3
4	4	5	3	5
5	5	4	5	4

Table 2.4 Mean and Overall Ranks of Five Graduate Students by Five Faculty Members on Industriousness

	Mean Ranking	*Overall Rank*
Student 1	1.75	2
Student 2	1.25	1
Student 3	3.25	3
Student 4	4.25	4
Student 5	4.5	5

Items in Interval-Level Categories. One can generate both the true interval-level boundaries of the response categories as well as the true interval level of each item's value using an approach called successive intervals (e.g., Ghiselli, Campbell, & Zedek, 1981). This is an indirect method, so it requires some conversions and, therefore, it is best to understand this by going through an example.

First, judges are required to place a number of stimuli (items) into a set number of categories (such as 5 or 10), where each category represents more or less of a particular dimension. For example, I may give 100 workers the following 10 stimuli and ask them to place the statements about their jobs into five different categories on the dimension of "satisfyingness," where 1 is the least satisfying and 5 is the most satisfying.

My job is . . .

1. disgusting
2. fun
3. underpaid
4. rewarding
5. delightful
6. challenging
7. enjoyable
8. revolting
9. interesting
10. meaningful

First, the proportions of the 100 workers who sorted the items into the five categories ranging from least favorable (1) to most favorable (5) are reported. These are shown in Table 2.5. Then a cumulative proportion matrix across the rows, leaving out the 5th column as it will always be 1.00 and does not provide any information, is created (see Table 2.6).

Next, proportions are changed to normal (z) scores, and the average of those z scores across the rows and down the columns are calculated. The column averages indicate the interval-level boundary for each category. Each row (item) average is then subtracted from the *grand mean* of the rows and columns (in our case, –0.17) and thus provides the interval-level stimulus value of each of our items. For example, item 1, "disgusting," has an interval-level value of (–0.17 – 3.41), which equals –3.58. Item 2, "fun," has an interval-level value of (–0.17 –1.68), which equals 1.51.

Note a couple of things in this matrix. The first is that the grand mean should be the same whether you generate it by taking the mean of the row averages or the column averages. This is a check on your calculations. Also, the sum of the scale scores for the items should be 0.00. In Table 2.7, these two conditions are met (within rounding error).

Table 2.5 Proportion of Judges Placing Each Statement Into One of Five Categories

	Category				
Item	*1 (least satisfying)*	*2*	*3*	*4*	*5 (most satisfying)*
disgusting	0.95	0.05	0.00	0.00	0.00
fun	0.00	0.00	0.50	0.40	0.10
underpaid	0.80	0.15	0.05	0.00	0.00
rewarding	0.00	0.00	0.50	0.25	0.25
delightful	0.00	0.00	0.05	0.05	0.90
challenging	0.00	0.10	0.60	0.20	0.10
enjoyable	0.00	0.10	0.40	0.35	0.15
revolting	0.90	0.10	0.00	0.00	0.00
interesting	0.00	0.05	0.30	0.50	0.15
meaningful	0.00	0.00	0.50	0.40	0.10

Table 2.6 Cumulative Proportion of Judges Placing Each Statement Into One of Five Categories

	Category				
Item	*1 (least satisfying)*	*2*	*3*	*4*	*5 (most satisfying)*
disgusting	0.95	1.00	1.00	1.00	1.00
fun	0.00	0.00	0.50	0.90	1.00
underpaid	0.80	0.95	1.00	1.00	1.00
rewarding	0.00	0.00	0.50	0.75	1.00
delightful	0.00	0.00	0.05	0.10	1.00
challenging	0.00	0.10	0.70	0.90	1.00
enjoyable	0.00	0.10	0.50	0.85	1.00
revolting	0.90	1.00	1.00	1.00	1.00
interesting	0.00	0.05	0.35	0.85	1.00
meaningful	0.00	0.00	0.50	0.90	1.00

Table 2.7 Normal Scores of Judges' Statements and Interval Scale and Item Boundary Values

	Category					
Item	*1*	*2*	*3*	*4*	*Row Average*	*Scale Score*
disgusting	1.65	4.00	4.00	4.00	3.41	–3.58
fun	–4.00	–4.00	0.00	1.28	–1.68	1.51
underpaid	1.28	1.65	4.00	4.00	2.73	–2.90
rewarding	–4.00	–4.00	0.00	0.68	–1.83	1.66
delightful	–4.00	–4.00	–1.65	–1.28	–2.73	2.56
challenging	–4.00	–1.28	0.52	1.28	–0.87	0.70
enjoyable	–4.00	–1.28	0.00	1.04	–1.06	0.89
revolting	1.28	4.00	4.00	4.00	3.32	–3.49
interesting	–4.00	–1.65	–0.39	1.04	–1.25	1.08
meaningful	–4.00	–4.00	0.00	1.28	–1.68	1.51
Column Average	–2.40	–1.06	1.05	1.73	**–0.17**	**- - - -**

From an interpretation perspective, the least to most satisfying items and their interval-level value can be reported:

1. disgusting (–3.58)
2. revolting (–3.49)
3. underpaid (–2.90)
4. challenging (0.70)
5. enjoyable (0.89)
6. interesting (1.08)
7./8. fun/meaningful (1.51 for both)
9. rewarding (1.66)
10. delightful (2.56)

Note that fun and meaningful have the same level of satisfyingness. So now that these items and the scale scores have been generated, what can be done with them? The items can now be administered to another sample, asking them if each of the items characterizes their jobs. Once their "yes" and "no" responses to each

item are obtained, the mean of the scores of the items to which they said "yes" will provide a measure of the desirableness of their jobs. For example, if I say "yes" to items 4, 5, and 6 but "no" to all the others, my job satisfaction would be $[(0.70 + 0.89 + 1.08)/3] = 0.89$. Once job satisfaction scores are generated this way for a sample of individuals, these interval-level scores can be used to (a) correlate with other variables, (b) compare jobs on their satisfaction levels, (c) compare industry satisfaction levels, and so forth.

A way to directly generate interval-level information about items is to use a method of equal-appearing intervals (Thurstone, 1929). To do so requires SMEs to make categorical judgments about a number of stimuli. It is absolutely essential that the item pool is large and encompasses the entire domain of interest. It is also important that the individuals making the judgments are SMEs. The strength of any conclusions will be based on the degree to which each of these criteria is met.

An example will be helpful here as well. Suppose I have 10 SMEs who are the job satisfaction gurus, and I ask them to take my 10 job characteristics and put them into five equal intervals. It is important that my SMEs assume that the intervals are indeed equal. So they do this, and the frequency with which each expert puts each item into each category is recorded. Then the means and standard deviations of these frequencies are calculated. This information is shown in Table 2.8.

In interpreting and using these findings, it can be seen that the "challenging" characteristic has the highest variability with a standard deviation of 1.03. This suggests that this item should perhaps be removed from the pool of items because it will

Table 2.8 Frequencies With Which Subject Matter Experts Placed Each Statement Into One of Five Equal-Interval Categories

	Category						
Item	*1*	*2*	*3*	*4*	*5*	*Row Mean*	*Row S.D.*
disgusting	8	2	0	0	0	1.2	0.42
fun	0	0	4	4	2	3.8	0.79
underpaid	6	3	1	0	0	1.5	0.71
rewarding	0	0	4	6	0	3.6	0.52
delightful	0	0	1	2	7	4.6	0.70
challenging	0	1	1	2	6	4.2	1.03
enjoyable	0	0	0	2	8	4.8	0.42
revolting	9	1	0	0	0	1.1	0.32
interesting	0	0	3	3	4	4.1	0.89
meaningful	0	0	0	7	3	4.3	0.48

have different meanings for different respondents in terms of how representative it is of job satisfaction. Another way to use this type of data is to ensure that items in the final scale that have widely differing levels of the attribute based on their mean scores are included. Therefore, if it is desirable to shorten the scale to three items, it would be appropriate to choose the items "disgusting" (mean of 1.2), "rewarding" (mean of 3.6), and "enjoyable" (mean of 4.8), as these three are likely to provide the highest range of scores from any subsequent sample of respondents.

Using a combination of the means and standard deviations, choosing between items becomes easier. Let's say that you needed to choose between the items "fun" and "rewarding." Because they are so close in their means (3.5 and 3.6), the standard deviation is examined and it shows that there was more consistency in SME responses on the "rewarding" item ($SD = 0.52$) than on the "fun" item ($SD = 0.79$). This would lead to the selection of "rewarding" over "fun" to represent the upper-middle range of satisfaction.

So, again, what can be done with such a scale? Similar to the indirect approach above, the mean score of each item that is retained in the item pool represents its scale value. Use the scale values to generate a score for each respondent. For example, if I say "yes," my job is characterized by being (a) underpaid (1.5), (b) rewarding (3.6), (c) challenging (4.2), and (d) meaningful (4.3) but not any of the other items, then my job satisfaction score is: $[(1.5 + 3.6 + 4.2 + 4.3)/4] = 3.4$. Once these satisfaction scores are generated for a sample, they can be used for other purposes.

Guttman Scales. A Guttman scale is another way of determining how the items are behaving in relation to one another. In the early history of experimental psychology, where psychophysical measures abounded, Guttman scaling made some sense. In the more amorphous realm of psychological constructs, however, the utility of Guttman scaling has all but disappeared. Interestingly, its theoretical background and rationale has resurfaced in modern test theory and so it is worthwhile to review the Guttman scaling approach.

In a Guttman scale, stimuli (otherwise known as test items) are presented in order of increasing extremeness. Knowing where the participants fall on the extremeness scale allows one to know what their responses were to all of the items—not just to the most extreme one. An example should help clarify this process. Let's assume we want to measure the construct of "veganness"—the degree to which an individual espouses being a vegan (someone who avoids using or consuming animal products). So we ask the following six questions; each is answered "yes" or "no":

1. I have restrictions on the type of food I eat.
2. I do not eat red meats.
3. I do not eat red meats or fowl.
4. I do not eat red meats, fowl, or fish.
5. I do not eat red meats, fowl, fish, or eggs.
6. I do not eat red meats, fowl, fish, eggs, or dairy products.

If someone responded to Item 6 with a "yes," he or she should have also have answered "yes" to Items 1–5. This individual would be on one extreme of the vegan scale. If an individual answered "no" to Item 1, he or she should also have answered "no" to Items 2–5. This person is at the other extreme of the vegan scale. The point where the individual shifts from saying "yes" to saying "no" determines his or her veganness.

The extent to which the items match a Guttman scale can be assessed by examining the pattern of responses. The more triangular the shape (like a staircase), the more Guttman-like the scale is acting. An example of seven people responding to the six vegan items is shown in Table 2.9. Person 7 is the most carnivorous and Person 1 is the most vegan. The others fall somewhere in between.

In the psychophysical realm, it is likely that the researcher is able to control the specific levels for a given stimuli, such as motion rates or luminance levels for visual stimuli or volume for auditory stimuli, and do so in a very fine-grained manner. Asking participants to respond with a "yes" or a "no" to "I see that" or "I hear that" will likely provide the researcher with a Guttman scale where increased steps in the stimuli correspond nicely to increased likelihood of participants seeing or hearing the stimuli. Those who see or hear the fainter stimuli are highly likely to see or hear the stronger stimuli.

Items for most psychological constructs, however, cannot easily be placed into a Guttman scale. For example, take the construct of leadership; how would one create items that would accurately reflect a simple staircase? What would a stimulus be that, if a person responded "yes" to it, would assure that that same person would have responded "yes" to all the other items representing less leadership? It simply does not make a lot of sense.

Sometimes individuals make the claim that items that form a Guttman scale response pattern make up a single construct (Guttman, 1947). However, this is not necessarily the case, even in a fairly straightforward example such as in an

Table 2.9 Guttman Scale Triangular Pattern for Seven Participants in Their Veganness

	Person						
Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	X						
2	X	X					
3	X	X	X				
4	X	X	X	X			
5	X	X	X	X	X		
6	X	X	X	X	X	X	

Note: X indicates a "yes" response.

achievement test. For example, suppose the following items are given to a sample ranging in age from 5 to 18:

1. What is the first letter in the alphabet?
2. What is the sum of 4 and 6?
3. What is the product of 7 and 9?
4. What is the capital of Great Britain?
5. What is the temperature for absolute zero?

It is very likely that the Guttman triangular form would fit responses to these questions perfectly. If a person gets the last item correct, he or she is likely to have gotten all of the rest correct. However, the items range from science to geography to math to verbal skills. In other words, responses that fall into a Guttman triangular form is not sufficient evidence to make the claim that the items form a single underlying construct.

The example above was created to demonstrate that it is important to consider the population or populations that will be responding to the scale, as well as the items, when developing a scale. A Guttman scalogram analysis would more likely proceed in the following manner. Let's use the example of being a team player in an organizational setting. There are four items in the scale of increasing extremeness in terms of whether or not moving to a team-based work environment would be beneficial to the organization. A sample of 10 workers is asked to respond to these items, and the data will be used to demonstrate that the team player construct is indeed a single construct. The items are as follows:

1. Teams might be helpful in this organization.
2. Teams would likely be helpful in this organization.
3. Teams would definitely be helpful in this organization.
4. Teams will be critical to the survival of this organization.

The pattern of responding "yes" to the items is shown in Table 2.10, with an X indicating agreement. It is anticipated that if individuals respond "yes" to item 4, they will have responded "yes" to the first three items as well. Conversely, if they do not respond "yes" to item 1, they will likely not respond "yes" to item 2. The number of times that unexpected responses, called reversal errors, occur provides an index of the degree of non-unidimensionality in the set of items.

Reversal errors can be used to assess the reproducibility or consistency of the scale items. The formula for calculating the *reproducibility coefficient* is

$$\text{Reproducibility} = [1 - (\text{total errors}/\text{total responses})] \times 100. \quad (2\text{–}1)$$

In this example, it is observed that Person D said "yes" to Items 3 and 4 but not to Items 1 and 2. Person I said "yes" to Items 2 and 3 but not to Item 1. Person J said

Table 2.10 Guttman Scalogram for 10 Respondents to Four Team Player Items

Person	*Item 1*	*Item 2*	*Item 3*	*Item 4*	*Errors*
A	X	X	X	X	0
B	X	X			0
C	X	X	X		0
D			X	X	2
E	X	X	X	X	0
F	X	X	X		0
G					0
H	X	X			0
I		X	X		1
J		X		X	2

Note: X indicates a "yes" response

"yes" to Items 2 and 4 but not to Items 1 and 3. This gives a total of 5 errors out of a total of 40 responses (each of 10 participants responded to 4 items). The reproducibility index then is

$$\begin{aligned}\text{Reproducibility} &= [1 - (5/40)] \times 100,\\ &= (1 - 0.125) \times 100,\\ &= 0.875 \times 100 = 87.5\%.\end{aligned}$$

Reproducibility indices less than 85% are low and indicate a need for items to be rewritten or deleted.

The difficulty in creating items that would generate a Guttman triangular form for the more amorphous psychological constructs (which are often of most interest) has made the Guttman scale generally not very practical. Furthermore, the notion of assessing the degree to which individuals possess an underlying trait (such as veganness or leadership) has moved forward with great strides in modern test theory, which leaves Guttman scaling somewhat obsolete.

Assessing Behaviors

How attitude measurement has developed over the past several decades has been covered in some depth. The assessment of behaviors rather than attitudes uses a somewhat different approach. In creating behavior-based scale items, the process is similar to that of any other scale development. The focus of the literature search and SME interviews would include clearly identifying the target behaviors and defining them in a way that is accepted. Because the focus of many behavioral scales

is their use in assessing behavioral change, it is critical to identify the antecedent conditions that trigger the behavior as well as the consequences of the behavior (both positive and negative).

Critical Incident Technique. A popular approach to developing items for behaviorally based scales is the critical incident technique (Flanagan, 1954). An illustration will be better than a verbal description, and the Behavioral Observation Scale (BOS) development (Latham & Wexley, 1977) provides a good example. BOSs are used in the context of rating employee performance. The critical incidents in the job performance area are those behaviors that are critical in determining good and poor performance. Over a period of time—say a month—supervisors carefully monitor the performance of the employees who report to them. The supervisors then rate the frequency with which the behaviors are demonstrated by the employees. For example, if the behavior to be rated is "greets the customer within 30 seconds of the customer entering the store," the supervisor would rate how often each employee does so over a fixed period of time.

After the allotted time period, ratings on each item are correlated to the total score across items. The ones with the highest correlations are presumed to be the most important for job performance and are retained for the final scale. As noted, the BOS is just one application of the critical incident technique, and it can be applied in a wide variety of settings.

Pilot Testing

There is no way to overemphasize the need to pilot test scale items. Asking colleagues, friends, family members, small groups of potential samples, and so forth to complete the scale is a critical step in the process of its development. Ask them to time how long it takes to complete the scale. Ask them to note where they have questions or clarity problems. Ask them to point out typos, grammatical errors, or anything else they can spot as they go through the scale. Ask them if items are too difficult (in achievement or ability tests).

A pilot test (or more than one pilot test) will provide feedback on all of these issues and serve an invaluable purpose. Issues raised by your pilot subjects can be taken care of before time and resources are wasted collecting useless data.

Summary and Next Step

In this chapter, the long process of creating a set of items and a scale to provide an index about an individual has been described. Specifically reviewed were

a. the empirical, theoretical, and rational approaches to item creation;

b. how to go about doing a thorough literature search and appropriate use of subject matter experts;

c. the principles in writing clear items;

d. the criteria to use in determining the appropriate number of items for a scale;

e. the history of creating attitude items;

f. what behavioral items should cover;

g. how to use the critical incident technique in creating behaviorally based items; and

h. the importance of pilot testing any scale.

The next step in creating a scale is to make several decisions about the types of responses that are to be obtained from those who answer the items on the scale. This is the topic of Chapter 3.

Problems and Exercises

1. An empirically developed set of test items are designed to do what?
2. Items developed theoretically have what characteristics?
3. Rationally developed tests use what techniques for item development and retention?
4. Using the construct you worked on for the Exercise in chapter 1, go to the existing literature and report how others have tried to develop the construct. What are their definitions? Use similar constructs to develop your own in more detail.
5. Rewrite each of the following items to improve them:
   a. My instructor presents material in an organized and enthusiastic manner.
   b. My instructor can be heard by everyone in the class.
   c. My instructor believes in what she or he is teaching.
   d. My instructor always provides feedback in a supportive manner.
   e. My instructor likes to use a variety of ways and materials to present information, such as videos, group discussions, projects, and outside speakers, so that the students have the opportunity to learn in a way most appropriate for their learning style.
6. Design five (5) items that will assess your construct of interest. Be sure to use SMEs, critical incidents, and existing information to generate some items.
7. Using the items you've designed, ask at least four colleagues to make paired comparisons of which item is most extreme in assessing your construct. With five items, you will have 10 paired comparisons. Set the data up as in Table 2.1, with your five items across the first row and down the first column. Put the proportion of people who rated the item as more extreme than the other item in the appropriate cell in the upper triangle of the table. Put 1 minus

this proportion in the opposing cell. Change the proportions to normalized *z* values as in Table 2.2, sum the columns, and then take the averages of those sums. Now you have a good idea as to the relative extremeness of your items in terms of how well they capture the construct.

8. Now ask another group of four colleagues to rank order the items from least (1) to most extreme (5). Record their rankings and then average the ranks. You now have another assessment of how extreme your items are in assessing the construct. Take your five items that you now have some confidence in regarding their rank-order of extremeness and put them in order of least extreme to most extreme. Now ask another group of four colleagues to respond "yes" or "no" to the items. Set up your data as in Table 2.9, with items in the first column, participants as the first row, and participant ratings of "yes" marked as an X in each cell. See if your scale follows a Guttman scalogram pattern. Calculate the reproducibility index for your scale.