SECOND EDITION

Psychological Testing

An Introduction

George Domino

University of Arizona

Marla L. Domino

Department of Mental Health, State of South Carolina

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521861816

© Cambridge University Press 2006

This publication is in copyright. Subject to statutory exception and to the provision of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published in print format 2006

ISBN-13 978-0-511-22012-8 eBook (EBL)
ISBN-10 0-511-22012-X eBook (EBL)

ISBN-13 978-0-521-86181-6 hardback
ISBN-10 0-521-86181-0 hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

6 Attitudes, Values, and Interests

AIM This chapter looks at the measurement of attitudes, values, and interests. These three areas share much in common from a psychometric as well as a theoretical point of view; in fact, some psychologists argue that the three areas, and especially attitudes and values, are not so different from each other. Some authors regard them as subsets of personality, while others point out that it is difficult, if not impossible, to define these three areas so that they are mutually exclusive.

The measurement of attitudes has been a central topic in social psychology, but has found relatively little application in the assessment of the individual client. Interest measurement on the other hand, particularly the assessment of career interests, probably represents one of the most successful applications of psychological testing to the individual client. The assessment of values has had somewhat of a mixed success, with such assessment often seen as part of personality and/or social psychology, and with some individual practitioners believing that values are an important facet of a client's assessment.

In the area of attitudes we look at some general issues, some classical ways of developing attitude scales, and some other examples to illustrate various aspects. In the area of values, we look at two of the more popular measures that have been developed, the Study of Values and the Rokeach Value Survey. Finally, in the area of interest measurement, we focus on career interests and the two sets of tests that have dominated this field, the Strong and the Kuder.

ATTITUDES

Definition. Once again, we find that there are many ways of defining attitudes and not all experts in this field agree as to what is and what is not an attitude. For our purposes however, we can consider attitudes as a predisposition to respond to a social object, such as a person, group, idea, physical object, etc., in particular situations; the predisposition interacts with other variables to influence the actual behavior of a person (Cardno, 1955).

Most discussions and/or definitions of attitude involve a tripartite model of affect, behavior, and cognition. That is, attitudes considered as a response to an object have an emotional component (how strongly one feels), a behavioral component (for example, voting for a candidate; shouting racial slurs; arguing about one's views), and a cognitive (thinking) component (e.g., Insko & Schopler, 1967; Krech, Crutchfield, & Ballachey, 1962). These three components should converge (that is, be highly similar), but each should also contribute something unique, and that indeed seems to be the case (e.g., Breckler, 1984; Ostrom, 1969; Rosenberg, Hovland, McGuire, et al., 1960). This tripartite model is the "classical" model that has guided much research, but it too has been criticized and new theoretical models proposed (e.g., Cacioppo, Petty, & Geen, 1989; Pratkanis & Greenwald, 1989; Zanna & Rempel, 1988).

Some writers seem to emphasize one component more than the others. For example, Thurstone (1946) defined attitude as, “the degree of positive or negative affect associated with some psychological object.” But most social scientists do perceive attitudes as learned predispositions to respond to a specific target, in either a positive or negative manner. As in other areas of assessment, there are a number of theoretical models available (e.g., Ajzen & Fishbein, 1980; Bentler & Speckart, 1979; Dohmen, Doll, & Feger, 1989; Fishbein, 1980; Jaccard, 1981; Triandis, 1980; G. Wiechmann & L. A. Wiechmann, 1973).

Centrality of attitudes. The study of attitudes and attitude change have occupied a central position in the social sciences, and particularly in social psychology, for a long time. Even today, the topic is one of the most active topics of study (Eagly & Chaiken, 1992; Oskamp, 1991; Rajecki, 1990). Part of the reason why the study of attitudes has been so central focuses on the assumption that attitudes will reveal behavior and because behavior seems so difficult to assess directly, attitudes are assumed to provide a way of understanding behavior (Kahle, 1984). Thus the relationship between attitudes and behavior is a major question, with some writers questioning such a relationship (e.g., Wicker, 1969) and others proposing that such a relationship is moderated by situational or personality factors (e.g., Ajzen & Fishbein, 1973; Zanna, Olson, & Fazio, 1980).

Some precautions. Henerson, Morris, and Fitz-Gibbon (1987) suggest that in the difficult task of measuring attitudes, we need to keep in mind four precautions:

1. Attitudes are inferred from a person’s words and actions; thus, they are not measured directly.
2. Attitudes are complex; feelings, beliefs, and behaviors do not always match.
3. Attitudes may not necessarily be stable, and so the establishment of reliability, especially when viewed as consistency over time, can be problematic.
4. Often we study attitudes without necessarily having uniform agreement as to their nature.

Ways of studying attitudes. There are many ways in which attitudes can be measured or assessed. The first and most obvious way to learn what a person’s attitude is toward a particular issue is to ask that person directly. Everyday conversations are filled with this type of assessment, as when we ask others such questions as “How do you feel about the death penalty?” “What do you think about abortion?” and “Where do you stand on gay rights?” This method of self-report is simple and direct, can be useful under some circumstances, but is quite limited from a psychometric point of view. There may be pressures to conform to majority opinion or to be less than candid about what one believes. There may be a confounding of expressed attitude with verbal skills, shyness, or other variables. A. L. Edwards (1957a) cites a study in which college students interviewed residents of Seattle about a pending legislative bill. Half of the residents were asked directly about their views, and half were given a secret and anonymous ballot to fill out. More “don’t know” responses were obtained by direct asking, and more unfavorable responses were obtained through the secret ballot. The results of the secret ballot were also in greater agreement with actual election results held several weeks later.

There are other self-reports, and these can include surveys, interviews, or more “personal” procedures such as keeping a log or journal. Self-reports can ordinarily be used when the respondents are able to understand what is being asked, can provide the necessary information, and are likely to respond honestly.

Observing directly. Another approach to the study of attitudes is to observe a person’s behavior, and to infer from that behavior the person’s attitudes. Thus, we might observe shoppers in a grocery store to determine their attitudes toward a particular product. The problem of course, is that a specific behavior may not be related to a particular attitude (for a brief, theoretical discussion of the relationship between attitudes and observable behavior see J. R. Eiser, 1987). You might buy chicken not because you love chicken but because you cannot afford filet mignon, or because you might want to try out a new recipe, or because your physician has suggested less red meat. Such *observer-reports* can include a variety of procedures ranging from observational assessment, to interviews, questionnaires, logs, etc. This approach is used when the people whose attitudes are being investigated may not be able to provide accurate information, or when the focus is directly on behavior that can be observed, or when there is evidence to suggest that an observer will be less biased and more objective.

Assessing directly. Because of the limitations inherent in both asking and observing, *attitude scales* have been developed as a third means of assessing attitudes. An attitude scale is essentially a collection of items, typically called *statements*, which elicit differential responses on the part of individuals who hold different attitudes. As with any other instrument, the attitude scale must be shown to have adequate reliability and validity. We will return to attitude scales below.

Sociometric procedures. Mention should be made here of *sociometric procedures*, which have been used to assess attitudes, not so much toward an external object, but more to assess the social patterns of a group. Thus, if we are interested in measuring the social climate of a classroom (which children play with which children; who are the leaders and the isolates, etc.), we might use a sociometric technique (for example, having each child identify their three best friends in that classroom). Such nominations may well reflect racial and other attitudes. Sociometric techniques can also be useful to obtain a *base rate* reading prior to the implementation of a program designed to change the group dynamics, or to determine whether a particular program has had an effect. There are a wide variety of sociometric measures, with two of the more popular consisting of *peer ratings* and *social choices.* In the peer rating method, the respondent reads a series of statements and indicates to whom the statement refers. For example:

___ this child is always happy.
___ this child has lots of friends.
___ this child is very good at playing sports.

In the social choice method, the respondent indicates the other persons whom he or she prefers. For example:

I would like to work with:_____
I would like to be on the same team as:_____

In general, it is recommended that sociometric items be positive rather than negative and general rather than specific (see Gronlund, 1959, for information on using and scoring sociometric instruments).

Records. Sometimes, *written records* that are kept for various purposes (e.g., school attendance records) can be analyzed to assess attitudes, such as attitudes toward school or a particular school subject.

Why use rating scales? Given so many ways of assessing attitudes, why should rating scales be used? There are at least six major reasons offered in the literature: (1) attitude rating scales can be administered to large groups of respondents at one sitting; (2) they can be administered under conditions of anonymity; (3) they allow the respondent to proceed at their own pace; (4) they present uniformity of procedure; (5) they allow for greater flexibility – for example, take-home questionnaires; and (6) the results are more amenable to statistical analyses.

At the same time, it should be recognized that their strengths are also their potential weaknesses. Their use with large groups can preclude obtaining individualized information or results that may suggest new avenues of questioning.

Ways of Measuring Attitudes

The method of equal-appearing intervals. This method, also known as the Thurstone method after its originator (Thurstone & Chave, 1929), is one of the most common methods of developing attitude scales and involves the following steps:

1. The first step is to select the social object or target to be evaluated. This might be an individual (the President), a group of people (artists), an idea or issue (physician-assisted suicide), a physical object (the new library building), or other targets.
2. Next a pool of items (close to 100 is not uncommon) is generated – designed to represent both favorable and unfavorable views. An assumption of most attitude research is that

attitudes reflect a bipolar continuum ranging from pro to con, from positive to negative.

3. The items are printed individually on cards, and these cards are then given to a group of "expert" subjects (judges) who individually sort the items into 11 piles according to the degree of favorableness (*not* according to whether they endorse the statement). Ordinarily, items placed in the first pile are the most unfavorable, items in the 6th pile are neutral, and items in the 11th pile are the most favorable. Note that this is very much like doing a Q sort, but the individual judge can place as many items in any one pile as he or she wishes. The judges are usually chosen because they are experts on the target being assessed – for example, statements for a religion attitude scale might be sorted by ministers.

4. The median value for each item is then computed by using the pile number. Thus if item #73 is placed by five judges in piles 6, 6, 7, 8, and 9, the median for that item would be 7. Ordinarily of course, we would be using a sizable sample of judges (closer to 100 is not uncommon), and so the median values would most likely be decimal numbers.

5. The median is a measure of central tendency – of average. We also need to compute for each item the amount of variability or of dispersion among scores, the scores again being the pile numbers. Ordinarily, we might think of computing the standard deviation, but Thurstone computed the *interquartile range*, known as *Q*. The interquartile range for an item is based on the difference between the pile values of the 25th and the 75th percentiles. This measure of dispersion in effect looks at the variability of the middle 50% of the values assigned by the judges to a particular item. A small *Q* value would indicate that most judges agreed in their placement of a statement, while a larger value would indicate greater disagreement. Often disagreement reflects a poorly written item that can be interpreted in various ways.

6. Items are then retained that (1) have a wide range of medians so that the entire continuum is represented and (2) that have the smallest *Q* values indicating placement agreement on the part of the judges.

7. The above steps will yield a scale of maybe 15 to 20 items that can then be administered to a sample of subjects with the instructions to check those items the respondent agrees with. The items are printed in random order. A person's score on the attitude scale is the median of the scale values of all the items endorsed.

For example, let's assume we have developed a scale to measure attitudes toward the topic of "psychological testing." Here are six representative items with their medians and *Q* values:

		Median	***Q* value**
1.	I would rather read about psychological testing than anything else	10.5	.68
14.	This topic makes you really appreciate the complexity of the human mind	8.3	3.19
19.	This is a highly interesting topic	6.7	.88
23.	Psychological testing is OK	4.8	.52
46.	This topic is very boring	2.1	.86
83.	This is the worst topic in psychology	1.3	.68

Note that item 14 would probably be eliminated because of its larger *Q* value. If the other items were retained and administered to a subject who endorses items 1, 19, and 23, then that person's score would be the median of 10.5, 6.7, and 4.8, which would be 6.7.

The intent of this method was to develop an interval scale, or possibly a ratio scale, but it is clear that the zero point (in this case the center of the distribution of items) is not a true zero. The title "method of equal-appearing intervals" suggests that the procedure results in an interval scale, but whether this is so has been questioned (e.g., Hevner, 1930; Petrie, 1969). Unidimensionality, hopefully, results from the writing of the initial pool of items, in that all of the items should be relevant to the target being assessed and from selecting items with small *Q* values.

There are a number of interesting questions that can be asked about the Thurstone procedure. For example, why use 11 categories? Why use the median rather than the mean? Could the judges rate each item rather than sort the items? In general, variations from the procedures originally used by Thurstone do not seem to make much difference (S. C. Webb, 1955).

One major concern is whether the attitudes of the judges who do the initial sorting influences how the items are sorted. At least some studies have suggested that the attitudes of the judges, even if extreme, can be held in abeyance with careful instructions, and do not influence the sorting of the items in a favorable-unfavorable continuum (e.g., Bruvold, 1975; Hinckley, 1932).

Another criticism made of Thurstone scales is that the same total score can be obtained by endorsing totally different items; one person may obtain a total score by endorsing one very favorable item or 9 or 10 unfavorable items that would add to the same total. This criticism is, of course, not unique to the Thurstone method. Note that when we construct a scale we ordinarily assume that there is a continuum we are assessing (intelligence, anxiety, psychopathology, liberal-conservative, etc.) and that we can locate the position of different individuals on this continuum as reflected by their test scores. We ordinarily don't care how those scores are composed – on a 100-item classroom test, it doesn't ordinarily matter which 10 items you miss, your raw score will still be 90. But one can argue that it *ought* to matter. Whether you miss the 10 most difficult items or the 10 easiest items probably says something about your level of knowledge or test-taking abilities, and whether you miss 10 items all on one topic vs. 10 items on 10 different topics might well be related to your breadth of knowledge.

Example of a Thurstone scale. J. H. Wright and Hicks (1966) attempted to develop a liberalism-conservatism scale using the Thurstone method. This dimension is a rather popular one, and several such scales exist (e.g., G. Hartmann, 1938; Hetzler, 1954; Kerr, 1952; G. D. Wilson & Patterson, 1968). The authors assembled 358 statements that were sorted into an 11-point continuum by 45 college students in an experimental psychology class (could these be considered experts?). From the pool of items, 23 were selected to represent the entire continuum and with the smallest SD (note that the original Thurstone method called for computing the interquartile range rather than the SD – but both are measures of variability), To validate the scale, it was administered to college students, members of Young Democrat and Young Republican organizations, with Democrats assumed to represent the liberal point of view and Republicans the conservative.

Below are representative items from the scale with the corresponding scale values:

1.	All old people should be taken care of by the government.	2.30
10.	Labor unions play an essential role in American democracy.	4.84
16.	The federal government should attempt to cut its annual spending.	7.45
23.	Isolation (complete) is the answer to our foreign policy.	10.50

Note that the dimension on which the items were sorted was liberal vs. conservative, rather than pro or con.

The authors report a corrected internal consistency coefficient (split-half) of +.79, and a Guttman reproducibility score of .87 (see following disscussion). The correlation between political affiliation and scale score was +.64, with Young Democrats having a mean score of 4.81 and Young Republicans a mean score of 5.93. These two means are not all that different, and one may question the initial assumption of the authors that democrats equal liberal and republicans equal conservative, and/or whether the scale really is valid. Note also that the authors chose contrasted groups, a legitimate procedure, but one may well wonder whether the scale would differentiate college students with different political persuasions who have chosen not to join campus political organizations. Finally, many of the items on the scale have become outmoded. Perhaps more than other measures, attitude scales have a short "shelf life," and rapidly become outdated in content, making longitudinal comparisons somewhat difficult.

The method of summated ratings. This method, also known as the Likert method after its originator (Likert, 1932), uses the following sequence of steps:

1. and 2. These are the same as in the Thurstone method, namely choosing a target concept and generating a pool of items.

3. The items are administered to a sample of subjects who indicate for each item whether they "strongly agree," "agree," "are undecided," "disagree," or "strongly disagree" (sometimes a word like "approve" is used instead of agree). Note that these subjects are not experts as in the Thurstone method; they are typically selected because they are available (introductory psychology students), or they represent the population that eventually will be assessed (e.g., registered Democrats).

4. A total score for each subject can be generated by assigning scores of 5, 4, 3, 2, and 1 to the above categories, and reversing the scoring for unfavorably worded items; the intent here is to be consistent, so that ordinarily higher scores represent a more favorable attitude.

5. An item analysis is then carried out by computing for each item a correlation between responses on that item and total scores on all the items (to be statistically correct, the total score should be for all the other items, so that the same item is not correlated with itself, but given a large number of items such overlap has minimal impact).

6. Individual items that correlate the highest with the total score are then retained for the final version of the scale. Note therefore that items could be retained that are heterogeneous in content, but correlate significantly with the total. Conversely, we could also carry out an item analysis using the method of item discrimination we discussed. Here we could identify the top 27% high scorers and the bottom 27% low scorers, and analyze for each item how these two groups responded to that item. Those items that show good discrimination between high and low scorers would be retained.

7. The final scale can then be administered to samples of subjects and their scores computed. Such scores will be highly relative in meaning – what is favorable or unfavorable depends upon the underlying distribution of scores.

Note should be made that some scales are called Likert scales simply because they use a 5-point response format, but may have been developed without using the Likert procedure, i.e., simply by the author putting together a set of items.

Are five response categories the best? To some degree psychological testing is affected by inertia and tradition. If the first or major researcher in one area uses a particular type of scale, quite often subsequent investigators also use the same type of scale, even when designing a new scale. But the issue of how many response categories are best – "best" judged by "user-friendly" aspects and by reliability and validity – has been investigated with mixed results (e.g., Komorita & Graham, 1965; Masters, 1974; Remmers & Ewart, 1941). Probably a safe conclusion here is that there does not seem to be an optimal number, but that five to seven categories seem to be better than fewer or more.

In terms of our fourfold classification of nominal, ordinal, interval, and ratio scales, Likert scales fall somewhere between ordinal and interval. On the one hand, by adding the arbitrary scores associated with each response option, we are acting as if the scale is an interval scale. But clearly the scores are arbitrary – why should the difference between "agree" and "strongly agree" be of the same numerical magnitude as the difference between "uncertain" and "agree"? And why should a response of "uncertain" be assigned a value of 3?

The above two methods are the most common ways of constructing attitude scales. Both are based upon what are called *psychophysical methods*, ways of assessing stimuli on the basis of their physical dimensions such as weight, but as determined psychologically (How heavy does this object feel?). Interested readers should see A. L. Edwards (1957a) for a discussion of these methods as related to attitude scale construction. How do the Thurstone and Likert procedures compare? For example, would a Thurstone scale of attitudes toward physician assisted suicide correlate with a Likert scale of the same target? Or what if we used the same pool of items and scored them first using the Thurstone method and then the Likert method – would the resulting sets of scores be highly related? In general, studies indicate that such scales typically correlate to a fair degree (in the range of .60 to .95). Likert scales typically show higher split-half or test-retest reliability than Thurstone scales. Likert scales are also easier to construct and use, which is why there are more of them available (see Roberts, Laughlin, & Wedell, 1999 for more complex aspects of this issue). We now turn to a number of other methods, which though important, have proven less common.

	Ratings						
	1	2	3	4	5	6	7
Group	(each of these would be defined using the seven statements)						
Hispanics							
American Indians							
Blacks							
Italians							
Russians							
etc.							

FIGURE 6–1. Example of a Bogardus Scale using multiple targets.

The Bogardus (1925) method. This method was developed in an attempt to measure attitudes toward different nationalities. Bogardus simply asked subjects to indicate whether they would admit members of a particular nationality or race to different degrees of social contact as defined by these seven categories:

1. close kinship by marriage
2. membership in one's club (or as close friends)
3. live on the same street as neighbor
4. employment in the same occupation (or work in same office)
5. citizenship in this country
6. visitor in this country
7. would exclude from this country

The scale forms a continuum of social distance, where at one end a person is willing to accept the target person in a very intimate relationship and at the other extreme would keep the target person as far away as possible. The instructions ask the subject to check those alternatives that reflect his or her reaction and not to react to the best or the worst members of the group that the respondent might have known. The score is simply the rank of the lowest (most intimate) item checked. If the group being assessed is a racial group, such as Blacks, then the resulting score is typically called a *racial distance quotient.* Note that multiple ratings could be obtained by having a bivariate table, with one dimension representing racial groups and the other dimension representing the seven categories. Figure 6.1 illustrates this.

The Bogardus approach is a methodology, but also a unique scale, as opposed to the Thurstone and Likert methods, which have yielded a wide variety of scales. Therefore, it is appropriate here to mention reliability and validity. Newcomb (1950) indicated that split-half reliability of the Bogardus scale typically reaches .90 or higher and that the validity is satisfactory. There have been a number of versions of the Bogardus scale; for example, Dodd (1935) developed an equal-interval version of this scale for use in the Far East, while Miller and Biggs (1958) developed a modified version for use with children. In general however, the Bogardus social distance approach has had limited impact, and its use nowadays seems to be rare.

Guttman scaling. This method is also known as *scalogram analysis* (Guttman, 1944). There is little difficulty in understanding the Bogardus social distance scale, and we can think of the Guttman method as an extension. We can easily visualize how close or far away a particular person might wish to keep from members of a racial group, even though we may not understand and/or condone racial prejudice. Ordinarily, we would expect that if a person welcomes a member of a different race into their own family, they would typically allow that person to work in the same office, and so on. The social distance scale is a univariate scale, almost by definition, where a person's position on that scale can be defined simply by the point where the person switches response mode. Suppose, for example, I have a mild case of racial bias against Venusian Pincos; I would allow them in this country as visitors or citizens, and would not really object to working with them, but I certainly would not want them as neighbors, or close friends, and would simply die if my daughter married one of them. My point of change is from item 4 to item 3; knowing that point of change, you could reproduce all my seven responses, assuming I did not reverse myself. This is in fact what Guttman scaling is all about. In developing a Guttman scale, a set of items that form a scalable continuum (such as social distance) is administered to a group of subjects, and the pattern of responses is analyzed to see if they fit the Guttman model. As an example, let's assume we have only three items: A (on marriage), B (on close friends), and C (on neighbor), each item requiring agreement or disagreement. Note that with the three items, we could theoretically obtain the following patterns of response:

	Item A (marriage)	**Item B (close friends)**	**Item C (neighbor)**
Response	Agree	Disagree	Disagree
	Agree	Agree	Disagree
Patterns:	*Agree	Agree	Agree
	*Disagree	Agree	Agree
	*Disagree	Disagree	Agree
	*Disagree	Disagree	Disagree
	Agree	Disagree	Agree
	Disagree	Agree	Disagree

In fact, the number of possible response patterns is 2^N where N is the number of items; in this case 2^3 equals $2 \times 2 \times 2$ or 8. If however, the items form a Guttman scale, there should be few if any reversals, and only the four response patterns marked by an $*$ should occur. The ideal number of response patterns then becomes $N + 1$, or 4 in this example. We can then compute what is called the *coefficient of reproducibility*, which is defined as:

$$1 - \frac{\text{total number of errors}}{\text{total number of responses}}$$

where errors are any deviation from the "ideal" pattern. If the reproducibility coefficient is .90 or above, then the scale is considered satisfactory. Although the matter seems fairly straightforward, there are a number of complicating issues that are beyond the scope of this book (e.g., A. L. Edwards, 1957a; Festinger, 1947; Green, 1954; Schuessler, 1961).

Guttman scales are not restricted to social distance, but could theoretically be developed to assess any variable. Let's assume I am working with an elderly population, perhaps female clients living in a nursing home, and I wish to assess their degree of independence as far as food preparation is concerned. I might develop a Guttman scale that might look like this:

This client is able to:

(a) plan and prepare a meal on her own

(b) plan and prepare a meal with some assistance

(c) prepare a meal but must be given the ingredients

(d) prepare a meal but needs assistance

(e) she not prepare a meal on her own

We can think of reproducibility as reflecting unidimensionality, and Guttman scales are thus unidimensional scales. Note however, that the method does not address the issue of equal intervals or the arbitrariness of the zero point; thus Guttman scales, despite their methodological sophistication, are not necessarily interval or ratio scales. The Guttman methodology has had more of an impact in terms of thinking about scale construction than in terms of actual, useful scales. Such scales do of course exist, but the majority assess variables that are behavioral in nature (such as the range of movement or physical skills a person possesses), rather than variables that are more "psychodynamic." There are a number of other procedures used to develop attitude scales, which, like the Guttman approach, are fairly complex both in theory and in statistical procedures (e.g., Banta, 1961; Coombs, 1950; Green, 1954; Hays & Borgatta, 1954; Lazarsfeld, 1950, 1954, 1959). In fact, there seems to be agreement that attitudes are multidimensional and that what is needed are more sophisticated techniques than the simple unidimensional approaches of Thurstone and Likert.

The Semantic Differential (SemD). The SemD was developed as a way of assessing word meaning but because this technique has been used quite frequently in the assessment of attitudes it

My ideal self

good ___ : ___ : ___ : ___ : ___ : ___ : ___ bad

small ___ : ___ : ___ : ___ : ___ : ___ : ___ large

beautiful ___ : ___ : ___ : ___ : ___ : ___ : ___ ugly

passive ___ : ___ : ___ : ___ : ___ : ___ : ___ active

sharp ___ : ___ : ___ : ___ : ___ : ___ : ___ dull

slow ___ : ___ : ___ : ___ : ___ : ___ : ___ fast

dirty ___ : ___ : ___ : ___ : ___ : ___ : ___ clean

etc.

FIGURE 6–2. Example of a Semantic Differential Scale.

can legitimately be considered here. The SemD is a method of observing and measuring the psychological meaning of things, usually concepts. We can communicate with one another because words and concepts have a shared meaning. If I say to you, "I have a dog," you know what a dog is. Yet that very word also has additional meanings that vary from person to person. One individual may think of dog as warm, cuddly, and friendly while another person may think of dog as smelly, fierce, and troublesome. There are thus at least two levels of meaning to words: the *denotative* or dictionary meaning, and the *connotative* or personal meaning. Osgood (Osgood, Suci, & Tannenbaum, 1957) developed the SemD to measure the connotative meanings of concepts as points in a semantic space. That space is three-dimensional, like a room in a house, and the dimensions, identified through factor analysis, are *evaluative* (e.g., good-bad), *potency* (e.g., strong-weak), and *activity* (fast-slow). Four additional factorial dimensions have been identified: density (e.g., numerous-sparse), orderliness (e.g., haphazard-systematic), reality (e.g., authentic-fake), and familiarity (e.g., commonplace-exceptional) (Bentler & LaVoie, 1972; LaVoie & Bentler, 1974).

The SemD then consists of a series of bipolar adjectives separated by a 7-point scale, on which the respondent rates a given concept. Figure 6.2 gives an example of a SemD.

How does one develop a SemD scale? There are basically two steps. The first step is to choose the concept(s) to be rated. These might be famous persons (e.g., Mother Theresa, Elton John), political concepts (socialism), psychiatric concepts (alcoholism), therapeutic concepts (my ideal self), cultural groups (Armenians), nonsense syllables, drawings, photographs, or whatever other stimuli would be appropriate to the area of investigation.

The second step is to select the bipolar adjectives that make up the SemD. We want the scale to be short, typically around 12 to 16 sets of bipolar adjectives, especially if we are asking each respondent to rate several concepts (e.g., rate the following cities: New York, Rome, Paris, Istanbul, Cairo, and Caracas). Which adjectives would we use? Bipolar adjectives are selected on the basis of two criteria: factor representativeness and relevance. Typical studies of the SemD have obtained the three factors indicated above, so we would select four or five bipolar adjectives representative of each factor; the loadings of each adjective pair on the various factor dimensions are given in various sources (e.g., Osgood, Suci, & Tannenbaum, 1957; Snider & Osgood, 1969). The second criterion of relevance is a bit more difficult to implement. If the concept of Teacher were being rated, one might wish to use bipolar pairs that are relevant to teaching behavior such as organized vs. disorganized, or concerned

Table 6–1. SemD ratings from one subject for five brands of beer

SemD Scales	Brand A	Brand B	Brand C	Brand D	Brand E
Pleasant-unpleasant	6	2	6	5	3
Ugly-beautiful	5	2	5	5	2
Sharp-flat	6	1	4	6	2
Salty-sweet	7	1	5	6	3
Happy-sad	5	3	5	7	1
Expensive-cheap	6	2	7	7	2
Mean	5.83	1.83	5.33	6.00	2.17

about students vs. not concerned (note that the "bipolar adjectives" need not be confined to one word). However, other bipolar pairs that on the surface may not seem highly relevant, such as heavy-light, ugly-beautiful, might in fact turn out to be quite relevant, in distinguishing between students who drop out vs. those who remain in school, for example.

In making up the SemD scale, about half of the bipolar adjectives would be listed in reverse order (as we did in Figure 6.2) to counteract response bias tendencies, so that not all left-hand terms would be positive. A 7-point scale is typically used, although between 3 and 11 spaces have been used in the literature; with children, a 5-point scale seems more appropriate.

Scoring the SemD. The SemD yields a surprising amount of data and a number of analyses are possible. The raw scores are simply the numbers 1 through 7 assigned as follows:

Good 7: 6: 5: 4: 3: 2: 1 Bad

The numbers do not appear on the respondent's protocol. Other numbers could be used, for example +3 to –3, but little if anything is gained and the arithmetic becomes more difficult.

If we are dealing with a single respondent, we can compare the semantic space directly. For example, Osgood and Luria (1954) analyzed a case of multiple personality (the famous "3 faces of Eve"), clearly showing that each personality perceived the world in rather drastically different terms, as evidenced by the ratings of such concepts as father, therapist, and myself.

Research projects and the assessment of attitudes usually involve a larger number of respondents, and various statistical analyses can be applied to the resulting data. Let's assume for example, we are studying attitudes toward various brands of beer. Table 6.1 shows the results from one subject who was asked to rate each of five brands:

For the sake of simplicity, let's assume that the six bipolar pairs are all evaluative items. A first step would be to compute and compare the means. Clearly brands A, C, and D are evaluated quite positively, while brands B and E are not. If the means were group averages, we could test for statistical significance perhaps using an ANOVA design. Note that in the SemD there are three sources of variation in the raw scores: differences between concepts, differences between scales (i.e., items), and differences between respondents. In addition we typically have three factors to contend with.

Distance-cluster analysis. If two brands of beer are close together in semantic space, that is rated equivalently, they are alike in "meaning" (for e.g., brands C and D in Table 6.1). If they are separated in semantic space they differ in meaning (e.g., brands D and E). What is needed is a measure of the distance between any two concepts. Correlation comes to mind, but for a variety of reasons, it is not suitable. What is used is the D statistic:

$$D_{ij} = \sqrt{\sum d_{ij}^2}$$

that is, the distance between any two concepts i and j equals the square root of the sum of the differences squared. For example, the distance between brand A and brand B in the above example equals:

$$(6-2)^2 + (5-2)^2 + (6-1)^2 + (7-1)^2 + (5-3)^2 + (6-2)^2 = 106$$

and $D = \sqrt{106}$ or 10.3

We can do the same for every pair of concepts. If we have n concepts (5 in our example), we will compute

$$\frac{n(n-2)}{2} D \text{ values.}$$

These D values can be written down in a matrix:

	Brand B	**C**	**D**	**E**
Brand A	10.30	3.00	2.65	9.06
B		8.89	10.44	3.16
C			3.16	8.19
D				9.95

Such a D matrix can be analyzed in several ways but the aim is the same: to seek how the concepts cluster together. The smaller the D value the closer in meaning are the concepts. Visually we can see that our five brands fall into two clusters: brands A, C, and D vs. brands B and E. Statistically we can use a variety of techniques including correlation and factor analysis (Osgood, Suci, & Tannenbaum, 1957) or more specific techniques (McQuitty, 1957; Nunnally, 1962).

Although three major factors are obtained in the typical study with the SemD, it is highly recommended that an investigator using the SemD check the resulting factor structure because there may be concept-scale interactions that affect such structure (Piotrowski, 1983; Sherry & Piotrowski, 1986). The evaluative factor seems to be quite consistent across samples, but the other two dimensions, potency and activity, are less consistent.

The SemD has found wide use in psychology, with both adults and children; DiVesta (1965) for example, provides a number of bipolar adjectives that can be used with children. An example of a SemD scale can be found in the study of Poresky, Hendrix, Mosier, et al., (1988) who developed the Companion Animal Semantic Differential to assess a respondent's perception of a childhood companion animal such as a pet dog. They used 18 bipolar sets of adjectives (bad-good, clean-dirty, cuddly-not cuddly) and obtained 164 responses from high-school, college, and graduate students. They used a 6-point scale to score each item, rather than the more standard 7-point. For the entire scale, the Cronbach alpha was .90 indicating substantial reliability. A factor analysis indicated four factors: (1) an evaluative factor (represented by such items as loving-not loving); (2) a factor related to the monetary value of the animal (e.g., valuable-worthless); (3) a factor related to affective value (kind-cruel); and (4) a factor related to the "size" of the animal (cuddly-not cuddly). When only the items that had substantial loadings were kept, the 18-item scale became a 9-item scale, and the four factors collapsed into one, namely an evaluative factor. Scores on the 9-item scale correlated .96 with scores on the 18-item scale. In case you're wondering of what use might such a scale be, you should know that there is a considerable body of literature and interest on the therapeutic effects of pet ownership on the elderly, the handicapped, coronary-care patients, and others.

One of the major concerns about the SemD is whether in fact the bipolar adjectives are bipolar – are the terms that anchor each scale truly opposite in meaning and equidistant from a true psychological midpoint? Results suggest that for some adjective pairs the assumption of bipolarity is not met (e.g., R. F. Green & Goldfried, 1965; Mann, Phillips, & Thompson, 1979; Schriesheim & Klich, 1991).

Checklists. One way to assess attitudes, particularly toward a large number of issues, is the *checklist* approach. As its name implies, this approach consists of a list of items (people, objects, issues, etc.) to which the respondent is asked to indicate their attitude in some way – by checking those items they endorse, selecting "favorable" or "unfavorable" for each item, indicating approval-disapproval, etc.

This is a simple and direct approach, and because all subjects are asked to respond to the same items, there is comparability of measurement. On the other hand, some argue that the presentation of a number of items can result in careless responding and hence lowered reliability and validity. In addition, the response categories typically used do not allow for degree of preference. (I may favor the death penalty and check that item in the list, but my convictions may not be very strong and might be easily dissuaded.)

An example of the checklist approach in the assessment of attitudes can be found in the work of G. D. Wilson and Patterson (1968) who developed the *conservatism* or C scale.

The C Scale

The liberal-conservative dimension has been studied quite extensively, both as it relates to political issues and voting behavior and a personality syndrome. Many investigators use terms like authoritarianism, dogmatism, or rigidity to refer to this dimension. Perhaps the major scale in this area has been the F (fascist) scale developed in a study called *The Authoritarian Personality* (Adorno et al., 1950). The F scale was for a time widely used, but also severely criticized for being open to acquiescence response set, poor phrasing, and other criticisms. Numerous attempts have been made, not only to develop revised F scales but also new scales based on the approach used with the F scale, as well as entirely different methodologies, such as that used in the C scale.

G. D. Wilson and Patterson (1968) decided that they would use a list of brief labels or "catchphrases" to measure "conservatism," defined as "resistance to change" and a preference for "safe, traditional, and conventional" behavior (G. D. Wilson, 1973). Theoretically, G. D. Wilson and Patterson (1968) identified conservatism as characterized by seven aspects that included religious fundamentalism, intolerance of minority groups, and insistence on strict rules and punishments. On the basis of these theoretical notions, they assembled a pool of 130 items chosen intuitively as reflective of these characteristics. They performed three item analyses (no details are given) and chose 50 items for the final scale. The respondent is asked which items "do you favor or believe in" and the response options are "yes, ?, no." For half of the items, a "yes" response indicates conservatism, and for half of the items a "no" response indicates conservatism. Examples of items (with their conservative response) are: the "death penalty (y)," "modern art (n)," "suicide (n)," "teenage drivers (n)," and "learning Latin (y)."

G. D. Wilson and Patterson (1968) reported a corrected split-half correlation coefficient of .94 based on 244 New Zealand subjects. They also present considerable validity data including age trends (older persons score higher), gender differences (females score slightly higher), differences between collegiate political groups, and between scientists and a conservative religious group.

In a subsequent study, Hartley and Holt (1971) used only the first half of the scale, but found additional validity evidence in various British groups; for example, psychology undergraduate students scored lowest, while male "headmasters" scored higher (female college of education students scored highest of all!). On the other hand, J. J. Ray (1971) administered the scale to Australian military recruits (all 20-year-old males) and found an alpha coefficient of +.63 and a preponderance of "yes" responses. He concluded that this scale was not suitable for random samples from the general population.

Bagley, Wilson, and Boshier (1970) translated the scale into Dutch and compared the responses of Dutch, British, and New Zealander subjects. A factor analysis indicated that for each of the three samples there was a "strong" general factor (however, it only accounted for 18.7 of the variance, or less), and the authors concluded that not only was there a "remarkable degree of cross-cultural stability" for the scale, but that the C scale had "considerable potential as an international test of social attitudes." The C scale was originally developed in New Zealand, and is relatively well known in English-speaking countries such as Australia, England, and New Zealand, but has found little utility in the United States. In part, this may be due to language differences (as Professor Higgins of My Fair Lady sings: English has not been spoken in the United States for quite some time!). For example, one C scale item is "birching" which means "paddling" as in corporal punishment administered by a teacher. In fact, a few investigators (e.g., Bahr & Chadwick, 1974; Joe, 1974; Joe & Kostyla, 1975) have adapted the C scale for American samples by making such item changes.

Although the reliability of the C scale would seem adequate (in the Dutch sample, the split-half was .89), Altemeyer (1981) brings up an interesting point. He argues that coefficient alpha, which you recall is one measure of reliability, reflects both the interitem correlations *and* the length of the test. Thus, one could have a questionnaire with a high coefficient alpha, but that might simply indicate that the questionnaire is long and not necessarily that the questionnaire is unidimensional. In fact, Altemeyer (1981) indicates that the average reliability coefficient for the C scale is .88, which indicates a mean interitem correlation of about .13 – thus, the C scale is criticized for not being unidimensional (see also Robertson & Cochrane, 1973).

Some general comments on rating scales. Like checklists, rating scales are used for a wide variety of assessment purposes, and the comments here, although they focus on attitude measurement, are meant to generalize to other areas of testing. Traditionally, rating scales were used to have one person assess another, for example, when a clinical psychologist might assess a client as to degree of depression, but the rating scales quickly were applied as self-report measures.

One common type of rating scale is *numerical* scale, where the choices offered to the respondent either explicitly or implicitly are defined numerically. For example, to the statement, "Suicide goes against the natural law," we might ask the respondent to indicate whether they (a) strongly agree, (b) agree, (c) are not sure, (d) disagree (e) strongly disagree. We may omit the numbers from the actual form seen by the respondent, but we would assign those numbers in scoring the response. Sometimes, the numbers are both positive and negative as in:

strongly agree	agree	not sure	disagree	strongly disagree
+2	+1	0	−1	−2

In general, such use of numbers makes life more complicated for both the respondent and the examiner. Mention should be made here, that there seems to be a general tendency on the part of some respondents to avoid extreme categories. Thus the 5-point scale illustrated above may turn out to be a 3-point scale for at least some subjects. The extension of this argument is that a 7-point scale is really preferable because in practice it will yield a 5-point scale.

Another type of rating scale is the *graphic* scale where the response options follow a straight line or some variation. For example:

How do you feel about capital punishment?
Place a check mark on the line:

---

1. should be abolished
2. should be used only for serious & repeat offenses
3. should be used for all serious offenses
4. is a deterrent & should be retained
5. should be used for all career criminals

Another example:
Where would you locate President Clinton on the following scale?

An excellent leader.

Better than most prior presidents.

Average in leadership.

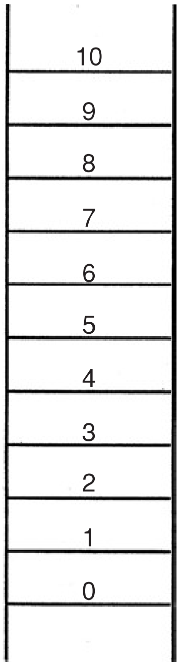
Less capable than most other presidents.

Totally lacking in leadership capabilities.

Note that a scale could combine both numerical and graphic properties; essentially what distinguishes a graphic scale is the presentation of some device, such as a line, where the respondent can place their answer. Note also, that from a psychometric point of view, it is easier to "force" the respondent to place their mark in a particular segment, rather than to allow free reign. In the capital punishment example above, we could place little vertical lines to distinguish and separate the five response options. Or we could allow the respondent to check anywhere on the scale, even between responses, and generate a score by actually measuring the distance where they placed their mark from the extreme left-hand beginning of the line. Guilford (1954) discusses these scales at length, as well as other less common types.

Self-anchoring scales. Kilpatrick and Cantril (1960) presented an approach that they called *self-anchoring* scaling, where the respondent is asked to describe the top and bottom anchoring points in terms of his or her own perceptions, values, attitudes, etc. This scaling method grew out of transactional theory that assumes that we live and operate in the world, through the self, both as personally perceived. That is, there is a unique reality for each of us – my perception of the world is not the same as your perception; what is perceived is inseparable from the perceiver.

Self-anchoring scales require both open-ended interviewing, content analysis, and nonverbal scaling. The first step is to ask the respondent to describe the "ideal" way of life. Second, he or she is asked to describe the "worst" way of life. Third, he or she is given a pictorial, nonverbal scale, such as an 11-point ladder:



The respondent is told that a 10 represents the ideal way of life as he or she described it, and 0 represents the worst way of life. So the two anchors have been defined by the respondent. Now the respondent is asked, "where on the ladder are you now?" Other questions may be asked, such as, "where on the ladder were you five years ago," "where will you be in two years," and so on.

The basic point of the ladder is that it provides a self-defined continuum that is anchored at either end in terms of personal perception. Other than that, the entire procedure is quite flexible. Fewer or more than 11 steps may be used; the numbers themselves may be omitted; a rather wide variety of concepts can be scaled; and instructions may be given in written form rather than as an interview, allowing the simultaneous assessment of a group of individuals.

Designing attitude scales. Oppenheim (1992), in discussing the design of "surveys," suggests a series of 14 steps. These are quite applicable to the design of attitude scales and are quite similar to the more generic steps suggested in Chapter 2. They are well worth repeating here (if you wish additional information on surveys, see Kerlinger, 1964; Kidder, Judd, & Smith, 1986; Rossi, Wright, & Anderson, 1983; Schuman & Kalton, 1985; Singer & Presser, 1989):

1. First decide the aims of the study. The aims should not be simply generic aims (I wish to study the attitudes of students toward physician-assisted suicide) but should be specific, and take the form of hypotheses to be tested (students who are highly authoritarian will endorse physician-assisted suicide to a greater degree than less authoritarian).
2. Review the relevant literature and carry out discussions with appropriate informants, individuals who by virtue of their expertise and/or community position are knowledgeable about the intended topic.
3. Develop a preliminary conceptualization of the study and revise it based on exploratory and/or in depth interviews.
4. Spell out the design of the study and assess its feasibility in terms of time, cost, staffing needed, and so on.
5. Spell out the *operational* definitions – that is, if our hypothesis is that "political attitudes are related to socioeconomic background," how will each of these variables be defined and measured?
6. Design or adapt the necessary research instruments.
7. Carry out pilot work to try out the instruments.
8. Develop a research design: How will respondents be selected? Is a control group needed? How will participation be ensured?
9. Select the sample(s).
10. Carry out the field work: interview subjects and/or administer questionnaires.
11. Process the data: code and/or score the responses, enter the data into the computer.
12. Carry out the appropriate statistical analyses.
13. Assemble the results.
14. Write the research report.

Writing items for attitude scales. Much of our earlier discussion on writing test items also applies here. Writing statements for any psychometric instrument is both an art and a science. A number of writers (e.g., A. L. Edwards, 1957a; A. L. Edwards & Kilpatrick, 1948; Payne, 1951; Thurstone & Chave, 1929; Wang, 1932), have made many valuable suggestions such as, make statements brief, unambiguous, simple, and direct; each statement should focus on only one idea; avoid double negatives; avoid "apple pie and motherhood" type of statements that everyone agrees with; don't use universals such as "always" or "never"; don't use emotionally laden words such as "adultery," "Communist," "agitator"; where possible, use positive rather than negative wording. For attitude scales, one difference is that factual statements, pertinent in achievement testing, do not make good items because individuals with different attitudes might well respond identically.

Ambiguous statements should not be used. For example, "It is important that we give Venusians the recognition they deserve" is a poor statement because it might be interpreted positively (Venusians should get more recognition) or negatively (Venusians deserve little recognition and that's what they should get). A. L. Edwards (1957a) suggested that a good first step in the preliminary evaluation of statements is to have a group of individuals answer the items first as if they had a favorable attitude and then as if they had an unfavorable attitude. Items that show a distinct shift in response are most likely useful items.

Closed vs. open response options. Most attitude scales presented in the literature use *closed* response options; this is the case in both the Thurstone and Likert methods where the respondent endorses (or not) a specific statement. We may also wish to use *open* response options, where respondents are asked to indicate in their own words what their attitude is – for example, "How valuable were the homework assignments in this class?" "Comment on the textbook used," and so on. Closed response options are advantageous from a statistical point of view. Open response options are more difficult to handle statistically, but can provide more information and allow respondents to express their feelings more directly. Both types of items can of course be used.

Measuring attitudes in specific situations. There are a number of situations where the assessment of attitudes might be helpful, but available scales may not quite fit the demands of the situation. For example, a city council may wish to determine how citizens feel toward the potential construction of a new park, or the regents of a university might wish to assess whether a new academic degree should be offered. The same steps we discussed in Chapter 2 might well be used here (or the steps offered by Oppenheim [1992] above). Perhaps it might not be necessary to have a "theory" about the proposed issue, but it certainly would be important to identify the objectives that are to be assessed and to produce items that follow the canons of good writing.

VALUES

Values also play a major role in life, especially because, as philosophers tell us, human beings are metaphysical animals searching for the purpose of their existence. Such purposes are guidelines for life or values (Grosze-Nipper & Rebel, 1987). Like the assessment of attitudes, the assessment of values is also a very complex undertaking, in part because values, like most other psychological variables, are constructs, i.e., abstract conceptions. Different social scientists have different conceptions and so perceive values differently, and there does not seem to be a uniformly accepted way of defining and conceptualizing values. As with attitudes, values cannot be measured directly, we can only infer a person's values by what they say and/or what they do. But people are complex and do not necessarily behave in logically consistent ways. Not every psychologist agrees that values are important; Mowrer (1967) for example, believed that the term "values" was essentially useless.

Formation and changes in values. Because of the central role that values occupy, there is a vast body of literature, both experimental and theoretical, on this topic. One intriguing question concerns how values are formed and how values change. Hoge and Bender (1974) suggested that there are three theoretical models that address this issue. The first model assumes that values are formed and changed by a vast array of events and experiences. We are all in the same "boat" and whatever affects that boat affects all of us. Thus, as our society becomes more violence-prone and materialistic, *we* become more violence-prone and materialistic. A second model assumes that certain developmental periods are crucial for the establishment of values. One such period is adolescence, and so high school and the beginning college years are "formative" years. This means that when there are relatively rapid social changes, different cohorts of individuals will have different values. The third model also assumes that values change developmentally, but the changes are primarily a function of age – for example, as people become older, they become more conservative.

The Study of Values (SoV)

The SoV (Allport, Vernon, & Lindzey, 1960; Vernon & Allport, 1931) was for many years the leading measure of values, used widely by social psychologists, in studies of personality, and even as a counseling and guidance tool. The SoV seems to be no longer popular, but it is still worthy of a close look. The SoV, originally published in 1931 and revised in 1951, was based on a theory (by Spranger, 1928) that assumed there were six basic values or personality types: theoretical, economic, aesthetic, social, political, and religious. As the authors indicated (Allport, Vernon, & Lindzey, 1960) Spranger held a rather positive view of human nature and did not consider the possibility of a "valueless" person, or someone who followed expediency (doing what is best for one's self) or hedonism (pleasure) as a way of life. Although the SoV was in some ways designed to operationalize Spranger's theory, the studies that were subsequently generated were only minimally related to Spranger's views; thus, while the SoV had quite an impact on psychological research, Spranger's theory did not.

The SoV was composed of two parts consisting of forced-choice items in which statements representing different values were presented, with the respondent having to choose one. Each of the 6 values was assessed by a total of 20 items, so the entire test was composed of 120 items. The SoV was designed primarily for college students or well-educated adults, and a somewhat unique aspect was that it could be hand scored by the subject.

Reliability. For a sample of 100 subjects, the corrected split-half reliabilities for the six scales range from .84 to .95, with a mean of .90. Test-retest reliabilities are also reported for two small samples, with a 1-month and a 2-month interval. These values are also quite acceptable, ranging from .77 to .93 (Allport, Vernon, & Lindzey, 1960). Hilton and Korn (1964) administered the SoV seven times to 30 college students over a 7-month period (in case you're wondering, the students were participating in a study of career decision making, and were paid for their participation). Reliability coefficients ranged from a low of .74 for the political value scale to a high of .91 for the aesthetic value scale. Subsequent studies have reported similar values.

An ipsative scale. The SoV is also an ipsative measure: if you score high on one scale you must score lower on some or all of the others. As the authors state in the test manual, it is not quite legitimate therefore to ask whether the scales intercorrelate. Nevertheless, they present the intercorrelations based on a sample of 100 males and a sample of 100 females. As expected, most of the correlations are negative, ranging in magnitude and sign from a −.48 (for religious vs. theoretical, in the female sample) to a +.27 for political vs. economic (in the male sample), and religious vs. social (in the female sample).

Validity. There are literally hundreds of studies in the literature that used the SoV, and most support its validity. One area in which the SoV has been used is to assess the changes in values that occur during the college years; in fact, K. A. Feldman and Newcomb (1969) after reviewing the available literature, believed that the SoV was the best single source of information about such changes. The study by Huntley (1965) although not necessarily representative, is illustrative and interesting. Huntley (1965), administered the SoV to male undergraduate college students at entrance to college and again just prior to graduation. Over a 6-year period some 1,800 students took the test, with 1,027 having both "entering" and "graduating" profiles. The students were grouped into nine major fields of study, such as science, engineering, and premed, according to their graduation status. Huntley (1965) then asked, and answered, four basic questions: (1) Do values (i.e., SoV scores) change significantly during the 4 years of college? Of the 54 possible changes (9 groups of students × 6 values), 27 showed statistically significant changes, with specific changes associated with specific majors. For example, both humanities and pre-med majors increased in their aesthetic value and decreased in their economic value, while industrial administration majors increased in both their aesthetic and economic values; (2) Do students who enter different majors show different values at entrance into college? Indeed they do. Engineering students, for example, have high economic and political values, while physics majors have low economic and political values; (3) What differences are found among the nine groups at graduation? Basically the same pattern of differences that exist at entrance. In fact, if the nine groups are ranked on each of the values, and the ranks at entrance are compared with those at graduation, there is a great deal of stability. In addition, what appears to happen is that value differences among groups are accentuated over the course of the four collegiate years; (4) Are there general trends? Considering these students as one cohort, theoretical, social, and political values show no appreciable change (keep in mind that these values *do* change for specific majors). Aesthetic values increase, and economic and religious values decrease, regardless of major.

Norms. The test manual presents norms based on 8,369 college students. The norms are subdivided by gender as well as by collegiate institution. In addition, norms are presented for a wide range of occupational groups, with the results supporting the construct validity of the SoV. For example, clergymen and theological students score highest on the religious value. Engineering students score highest on the theoretical value, while business administration students score highest on the economic and political scales. Subsequent norms included a national sample of high-school students tested in 1968, and composed of more than 5000 males and 7,000 females. Again, given the ipsative nature of this scale, we may question the appropriateness of norms.

Criticisms. Over the years, a variety of criticisms have been leveled at the SoV. For example, Gage (1959) felt that the SoV confounded interests and values. Others repeatedly pointed out that the values assessed were based on "ideal" types and did not necessarily match reality; furthermore, these values appeared to be closely tied to "middle class" values.

The Rokeach Value Survey (RVS)

Introduction. One of the most widely used surveys of values is the *Rokeach Value Survey* (RVS). Rokeach (1973) defined values as beliefs concerning either desirable *modes of conduct* or desirable *end-states of existence*. The first type of values is what Rokeach labeled *instrumental* values, in that they are concerned with modes of conduct; the second type of values are *terminal* values in that they are concerned with end states. Furthermore, Rokeach (1973) divided instrumental values into two types: moral values that have an interpersonal focus, and competence or self-actualization values that have a personal focus. Terminal values are also of two types: self-centered or personal, and society-centered or social.

Rokeach (1973) distinguished values from attitudes in that a value refers to a single belief, while an attitude concerns an organization of several beliefs centered on a specific target. Furthermore, values transcend the specific target, represent a standard, are much smaller in number than attitudes, and occupy a more central position in a person's psychological functioning.

Description. The RVS is a rather simple affair that consists of two lists of 18 values each, which the respondent places in rank order, in order of importance as guiding principles of their life. Table 6.2 illustrates the RVS. Note that each value is accompanied by a short, defining phrase.

Originally the RVS consisted simple of printed lists; subsequently, each value is printed on a removable gummed label, and the labels are placed in rank order. The two types of values, instrumental and terminal, are ranked and analyzed separately, but the subtypes (such as personal and social) are not considered. The RVS is then a self-report instrument, group administered, with no time limit, and designed for adolescents and adults. Rokeach (1973) suggests that the RVS is really a projective test, like the Rorschach Inkblot technique, in that the respondent has no

Table 6–2. RVS values

Terminal values	Instrumental values
A comfortable life (a prosperous life)	Ambitious (hard-working, aspiring)
An exciting life (a stimulating, active life)	Broadminded (open-minded)
A sense of accomplishment (lasting contribution)	Capable (competent, effective)
A world at peace (free of war and conflict)	Cheerful* (lighthearted, joyful)
A world of beauty (beauty of nature and the arts)	Clean (neat, tidy)
Equality (brotherhood, equal opportunity for all)	Courageous (standing up for your beliefs)
Family security (taking care of loved ones)	Forgiving (willing to pardon others)
Freedom (independence, free choice)	Helpful (working for the welfare of others)
Happiness* (contentedness)	Honest (sincere, truthful)
Inner harmony (freedom from inner conflict)	Imaginative (daring, creative)
Mature love (sexual and spiritual intimacy)	Independent (self-reliant, self-sufficient)
National security (protection from attack)	Intellectual (intelligent, reflective)
Pleasure (an enjoyable, leisurely life)	Logical (consistent, rational)
Salvation (saved, eternal life)	Loving (affectionate, tender)
Self-respect (self-esteem)	Obedient (dutiful, respectful)
Social recognition (respect, admiration)	Polite (courteous, well mannered)
True friendship (close companionship)	Responsible (dependable, reliable)
Wisdom (a mature understanding of life)	Self-controlled (restrained, self-disciplined)

**Note:* These values were later replaced by health and loyal respectively.
Adapted with the permission of The Free Press, a Division of Simon & Schuster from *The Nature Of Human Values* by Milton Rokeach. Copyright © 1973 by The Free Press.

guidelines for responding other than his or her own internalized system of values.

How did Rokeach arrive at these particular 36 values? Basically through a clinical process that began with amassing a large number of value labels from various sources (the instrumental values actually began as personality traits), eliminating those that were synonymous, and in some cases those that intercorrelated highly. Thus there is the basic question of content validity and Rokeach (1973) himself admits that his procedure is "intuitive" and his results differ from those that might have been obtained by other researchers.

Scoring the RVS. Basically, there is no scoring procedure with the RVS. Once the respondent has provided the two sets of 18 ranks, the ranks cannot of course be added together to get a sum because every respondent would obtain exactly the same score.

For a group of individuals we can compute for each value the mean or median of the rank assigned to that value. We can then convert these average values into ranks. For example, J. Andrews (1973) administered the RVS to 61 college students, together with a questionnaire to assess the degree of "ego identity" achieved by each student. Students classified as "high identity achievement" ranked the RVS instrumental values as follows:

value	mean ranking
honest	5.50
responsible	5.68
loving	5.96
broadminded	6.56
independent	7.48
capable	7.88
etc.	

We can change the mean rank values back to ranks by calling honest = 1, responsible = 2, loving = 3, and so on.

Another scoring approach would be to summate together subsets of values that on the basis of either a statistical criterion such as factor analysis, or a clinical judgment such as content analysis, seem to go together. For example, Silverman, Bishop, and Jaffe (1976) studied the RVS responses of some 954 psychology graduate students. To determine whether there were differences between students who studied different fields of psychology (e.g., clinical, experimental, developmental), the investigators computed the average of the median rankings assigned to "mature love," "true friendship," "cheerful," "helpful," and "loving" – this cluster of values was labeled "interpersonal affective values." A similar index called "cognitive competency" was calculated by averaging the median rankings for "intellectual" and "logical."

Reliability. There are at least two ways of assessing the temporal stability (i.e., test-retest reliability) of the RVS. One way is to administer the RVS to a group of individuals and retest them later. For each person, we can correlate the two sets of ranks and then can compute the median of such rank order correlation coefficients for our sample of subjects. Rokeach (1973) reports such medians as ranging from .76 to .80 for terminal values and .65 to .72 for instrumental values, with samples of college students retested after 3 weeks to 4 months.

Another way is also to administer the RVS twice, but to focus on each value separately. We may for example, start out with "a comfortable life." For each subject in our sample, we have the two ranks assigned to this value. We can then compute a correlation coefficient across subjects for that specific value. When this is done, separately for each of the 36 values, we find that the reliabilities are quite low; for the terminal values the average reliability is about .65 (Rokeach, 1973) and for the instrumental values it is about .56 (Feather, 1975). This is of course not surprising because each "scale" is made up of only one item. One important implication of such low reliability is that the RVS should *not* be used for individual counseling and assessment.

One problem, then, is that the reliability of the RVS is marginal at best. Rokeach (1973) presents the results of various studies, primarily with college students, and with various test-retest intervals ranging from 3 weeks to 16 months; of the 29 coefficients given, 14 are below .70, and all range from .53 to .87, with a median of .70. Inglehart (1985), on the other hand, looked at the results of a national sample, one assessed in 1968 and again in 1981. Because there were different subjects, it is not possible to compute correlation coefficients, but Inglehart (1985) reported that the stability of rankings over the 13-year period was "phenomenal." The six highest- and six lowest-ranked values in 1968 were also the six highest-and six lowest-ranked values in 1981.

It is interesting to note that all of the 36 values are socially desirable, and that respondents often indicate that the ranking task is a difficult one and they have "little confidence" that they have done so in a reliable manner.

Validity. Rokeach's (1973) book is replete with various analyses and comparisons of RVS rankings, including cross-cultural comparisons and analyses of such variables such as as race, socioeconomic status, educational level, and occupation. The RVS has also been used in hundreds of studies across a wide spectrum of topics, with most studies showing encouraging results that support the construct validity of this instrument. These studies range from comparisons of women who prefer "Ivory" as a washing machine detergent to studies of hippies (Rokeach, 1973). One area where the study of values has found substantial application is that of psychotherapy, where the values of patients and of therapists and their concomitant changes, have been studied (e.g., Beutler, Arizmendi, Crago, et al., 1983; Beutler, Crago, & Arizmendi, 1986; Jensen & Bergin, 1988; Kelly, 1990).

Cross-cultural aspects. Rokeach (1973) believed that the RVS could be used cross-culturally because the values listed are universal and problems of translation can be surmounted. On the other hand, it can be argued that these values are relevant to Western cultures only; for example, "filial piety," a central value for Chinese is not included in the RVS. It can also be argued that although the same word can be found in two languages, it does not necessarily have the same layers of meaning in the two cultures. Nevertheless, a number of investigators have applied the RVS cross-culturally, both in English-speaking countries such as Australia and non-Western cultures such as China (e.g., Feather, 1986; Lau, 1988; Ng et al., 1982).

An example of a cross-cultural application is found in the study by Domino and Acosta (1987), who administered the RVS to a sample of first generation Mexican Americans. These individuals were identified as being either "highly acculturated," that is more American, or "less acculturated," that is more Mexican. Their rankings of the RVS were then analyzed in various ways, including comparisons with the national norms

Table 6–3. Factor structure of the RVS Based on a sample of 1,409 respondents (Rokeach, 1973)

	Example of item with		
Factor	Positive loading	Negative loading	Percentage of variance
1. Immediate vs. delayed gratification	A comfortable life	Wisdom	8.2
2. Competence vs. religious morality	Logical	Forgiving	7.8
3. Self-constriction vs. self-expansion	Obedient	Broadminded	5.5
4. Social vs. personal orientation	A world at peace	True friendship	5.4
5. Societal vs. family security	A world of beauty	Family security	5.0
6. Respect vs. love	Social recognition	Mature Love	4.9
7. Inner vs. other directed	Polite	Courageous	4.0

provided by Rokeach and with local norms based on Anglos. These researchers found a greater correspondence of values between high acculturation subjects and the comparison groups than between the low acculturation subjects and the comparison groups – those that were more “American” in their language and general cultural identification were also more American in their values.

Factor analysis. Factor analytic studies do seem to support the terminal-instrumental differentiation, although not everyone agrees (e.g., Crosby, Bitner, & Gill, 1990; Feather & Peay, 1975; Heath & Fogel, 1978; Vinson et al., 1977). Factor analyses suggest that the 36 values are not independent of each other and that certain values do cluster together. Rokeach (1973) suggests that there are seven basic factors that cut across the terminal-instrumental distinction. These factors are indicated in Table 6.3. One question that can be asked of the results of a factor analysis is how “important” each factor is. Different respondents give different answers (ranks) to different values. This variation of response can be called “total variance.” When we identify a factor, we can ask how much of the total variance does that factor account for? For the RVS data reported in Table 6.3, factor 1 accounts for only 8.2% of the total variation, and in fact all seven factors together account for only 40.8% of the total variation, leaving 59.2% of the variation unaccounted for. This suggests that the factors are probably not very powerful, either in predicting behavior or in helping us to conceptualize values. Heath and Fogel (1978) had subjects rate rather than rank the importance of each of the 36 values; their results suggested eight factors rather than seven.

Norms. Rokeach (1973) presents the rankings for a group of 665 males and a group of 744 females, and these are presented in Table 6.4. Note that of the 36 values, 20 show significant gender differences. Even though the ranks may be identical, there may be a significant difference on the actual rank value assigned. The differences seem to be in line with the different ways that men and women are socialized in Western cultures, with males endorsing more achievement and intellectually oriented values, more materialistic and pleasure seeking, while women rank higher religious values, love, personal happiness, and lack of both inner and outer conflict.

Rank order correlation coefficient. Despite the caveat that the RVS should not be used for individual counseling, we use a fictitious example to illustrate the rank order correlation coefficient, designed to compare two sets of ranks. Let’s say that you and your fiance are contemplating marriage, and you wonder whether your values are compatible. You both independently rank order the RVS items. The results for the instrumental values are shown in Table 6.5. The question here is how similar are the two sets of values? We can easily calculate the rank order correlation coefficient (ρ) using the formula:

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

where N stands for the number of items being ranked; in this case $N = 18$. All we need to do is calculate for each set of ranks the difference

Table 6–4. Values medians and composite rank orders for American men and women (Rokeach, 1973)

Terminal value:	Male (n = 665)		Female (n = 744)		Lower rank shown by
A comfortable life	7.8	(4)	10.0	(13)	Males
An exciting life	14.6	(18)	15.8	(18)	Males
A sense of accomplishment	8.3	(7)	9.4	(10)	Males
A world at peace	3.8	(1)	3.0	(1)	Females
A world of beauty	13.6	(15)	13.5	(15)	–
Equality	8.9	(9)	8.3	(8)	–
Family security	3.8	(2)	3.8	(2)	–
Freedom	4.9	(3)	6.1	(3)	Males
Happiness	7.9	(5)	7.4	(5)	Females
Inner harmony	11.1	(13)	9.8	(12)	Females
Mature love	12.6	(14)	12.3	(14)	–
National security	9.2	(10)	9.8	(11)	–
Pleasure	14.1	(17)	15.0	(16)	Males
Salvation	9.9	(12)	7.3	(4)	Females
Self-respect	8.2	(6)	7.4	(6)	Females
Social recognition	13.8	(16)	15.0	(17)	Males
True friendship	9.6	(11)	9.1	(9)	–
Wisdom	8.5	(8)	7.7	(7)	Females
Instrumental values					
Ambitious	5.6	(2)	7.4	(4)	Males
Broadminded	7.2	(4)	7.7	(5)	–
Capable	8.9	(8)	10.1	(12)	Males
Cheerful	10.4	(12)	(9.4)	(10)	Females
Clean	9.4	(9)	8.1	(8)	Females
Courageous	7.5	(5)	8.1	(6)	–
Forgiving	8.2	(6)	6.4	(2)	Females
Helpful	8.3	(7)	8.1	(7)	–
Honest	3.4	(1)	3.2	(1)	–
Imaginative	14.3	(18)	16.1	(18)	Males
Independent	10.2	(11)	10.7	(14)	–
Intellectual	12.8	(15)	13.2	(16)	–
Logical	13.5	(16)	14.7	(17)	–
Loving	10.9	(14)	8.6	(9)	Females
Obedient	13.5	(17)	13.1	(15)	–
Polite	10.9	(13)	10.7	(13)	–
Responsible	6.6	(3)	6.8	(3)	–
Self-controlled	9.7	(10).	9.5	(11)	–

Note: The figures shown are median rankings and in parentheses composite rank orders.
The gender differences are based on median rankings.
Adapted with the permission of The Free Press, a Division of Simon & Schuster from *The Nature of Human Values* by Milton Rokeach. Copyright © 1973 by The Free Press.

between ranks, square each difference, and find the sum. This is done in Table 6.5, in the columns labeled D (difference) and D^2. The sum is 746, and substituting in the formula gives us:

$$= 1 - \frac{6\,(746)}{5814} = 1 - \frac{746}{969} = 1 - .77 = +.23$$

These results would suggest that there is a very low degree of agreement between you and your fiance as to what values are important in life, and indeed a perusal of the rankings suggest some highly significant discrepancies (e.g., self-controlled and courageous), some less significant discrepancies (e.g., cheerful and clean), and some near unanimity (ambitious and broadminded). If these results were reliable, one might predict some conflict ahead, unless of course you believe in the "opposites attract" school of thought

Table 6–5. Computational example of the rank order correlation coefficient using RVS data

Instrumental value	Your rank	Your fiance's rank	*D*	D^2
Ambitious	2	1	1	1
Broadminded	8	9	1	1
Capable	4	2	2	4
Cheerful	12	7	5	25
Clean	15	10	5	25
Courageous	5	16	11	121
Forgiving	6	12	6	36
Helpful	7	17	10	100
Honest	1	11	10	100
Imaginative	18	14	4	16
Independent	11	3	8	64
Intellectual	10	15	5	25
Logical	9	4	5	25
Loving	13	8	5	25
Obedient	14	18	4	16
Polite	16	13	3	9
Responsible	3	6	3	9
Self-controlled	17	5	12	144
				$\sum = 746$

rather than the "birds of a feather flock together" approach.

Criticisms. The RVS has been criticized for a number of reasons (Braithwaite & Law, 1985; Feather, 1975). It is of course an ipsative measure and yields only ordinal data; strictly speaking, its data should not be used with analysis of variance or other statistical procedures that require a normal distribution, although such procedures are indeed "robust" and seem to apply even when the assumptions are violated. Others have questioned whether the RVS measures what one prefers or what one *ought* to prefer (Bolt, 1978) and the distinction between terminal and instrumental values (Heath & Fogel, 1978).

One major criticism is that the rank ordering procedure does not allow for the assessment of intensity, which is basically the same criticism that this is not an interval scale. Thus two individuals can select the same value as their first choice, and only one may feel quite sanguine about it. Similarly, you may give a value a rank of 2 because it really differs from your number 1 choice, but the difference may be minimal for another person with the identical rankings. In fact, several researchers have modified the RVS into an interval measure (e.g., Moore, 1975; Penner, Homant, & Rokeach, 1968; Rankin & Grobe, 1980). Interestingly enough, some of the results suggest that rank-order scaling is a better technique than other approaches (e.g., Miethe, 1985).

INTERESTS

We now turn to the third area of measurement for this chapter, and that is interests, and more specifically, career interests. How can career interests be assessed? The most obvious and direct method is to ask individuals what they are interested in. These are called *expressed* interests, and perhaps not surprisingly, this is a reasonably valid method. On the other hand, people are often not sure what their interests are, or are unable to specify them objectively, or may have little awareness of how their particular interests and the demands of the world of work might dovetail. A second way is the assessment of such likes and dislikes through inventories. This method is perhaps the most popular method and has a number of advantages, including the fact that it permits an individual to compare their interests with those of other people, and more specifically with people in various occupations. A third way is to assume that someone interested in a particular occupation will have a fair amount of knowledge about that occupation, even before entering the occupation. Thus we could put together a test of knowledge about being a lawyer and assume that those who score high may be potential lawyers. That of course is a major assumption, not necessarily reflective of the real world. Finally, we can observe a person's behavior. If Johnny, a high school student, spends all of his spare time repairing automobiles, we might speculate that he is headed for a career as auto mechanic – but of course, our speculations may be quite incorrect.

The field of career interest measurement has been dominated by the work of two individuals. In 1927, E. K. Strong, Jr. published the Strong Vocational Interest Blank for Men, an empirically based inventory that compared a person's likes and dislikes with those of individuals in different occupations. The SVIB and its revisions became extremely popular and were used frequently in both college settings and private practice (Zytowski & Warman, 1982). In 1934, G. F. Kuder developed the Kuder Preference Record, which initially used content scales (e.g., agriculture) rather than specific occupational scales. This test also proved quite popular and underwent a number of revisions.

A third key event in the history of career interest assessment occurred in 1959, when John Holland published a theory regarding human behavior that found wide applicability to career interest assessment. Holland argued that the choice of an occupation is basically a reflection of one's personality, and so career-interest inventories are basically personality inventories.

Much of the literature and efforts in career assessment depend on a general assumption that people with similar interests tend to enter the same occupation, and to the degree that one's interests are congruent with those of people in that occupation, the result will be greater job satisfaction. There certainly seems to be substantial support for the first part of that assumption, but relatively little for the second part.

The Strong Interest Inventory (SII)

Introduction. The Strong Vocational Interest Blank for Men (SVIB) is the granddaddy of all career-interest inventories, developed by E. K. Strong, and originally published in 1927. A separate form for women was developed in 1933. The male and female forms were each revised twice, separately. In 1974, the two gender forms were merged into one. The SVIB became the Strong-Campbell Interest Inventory (SCII) and underwent extensive revisions (D. P. Campbell, 1974; D. P. Campbell & J. C. Hansen, 1981; J. C. Hansen & D. P. Campbell, 1985), including the development of occupational scales that were traditionally linked with the opposite sex. For example, a nursing scale for males and a carpenter and electrician scales for women. Recently, the name was changed to the Strong Interest Inventory (SII) (or Strong for short), and a 1994 revision published. To minimize confusion and reduce the alphabet soup, the word *Strong* is used to refer to any of these inventories (except in the rare instances where this would violate the intended meaning).

Description. Basically, the Strong compares a person's career interests with those of people who are satisfactorily employed in a wide variety of occupations. It is thus a measure of interests, not of ability or competence. The Strong contains 325 items grouped into seven sections. The bulk of the items (first five sections) require the respondent to indicate like, dislike, or indifferent to 131 occupations (Would you like to be a dentist? a psychologist?), 36 school subjects (algebra, literature), 51 career-related activities (carpentry; gardening; fund raising), 39 leisure activities (camping trips; cooking), and 24 types of people (Would you like to work with children? the elderly? artists?). Section 6 requires the respondent to select from pairs of activities that they prefer (Would you prefer working with "things" or with people?), and section 7 has some self-descriptive statements (Are you a patient person?). Strong originally used these various types of items in an empirical effort to see which type worked best. Subsequent research suggests that item content is more important than item format, and so the varied items have been retained also because they relieve the monotony of responding to a long list of similar questions (D. P. Campbell, 1974).

The primary aim of the Strong is for counseling high school and college students as well as and adults who are college graduates, about their career choices. It and particularly focuses on those careers that attract college graduates, rather than blue-collar occupations or skilled trades such electrician and plumber. Thus the Strong is geared primarily for age 17 and older. Career interests seem to stabilize for most people between the ages of 20 and 25, so the Strong is most accurate for this age range; it does not seem to be appropriate or useful for anyone younger than 16.

It is not the intent of the Strong to tell a person what career they should enter or where they can be successful in the world of work. In fact, the Strong has little to do with competence and capabilities; a person may have a great deal of similarity of interest with those shown by physicians, but have neither the cognitive abilities nor the educational credentials required to enter and do well in medical school.

There are at least two manuals available for the professional user: the Manual, which contains the technical data (J. C. Hansen & D. P. Campbell, 1985), and the User's Guide (J. C. Hansen, 1984), which is more "user friendly" and more of a typical manual.

Item selection. Where did the items in the Strong come from? Originally, they were generated by Strong and others, and were basically the result of "clinical insight." Subsequently, the items contained in the current Strong came from earlier editions and were selected on the basis of their psychometric properties (i.e., reliability and validity), as well as on their "public relations" aspects – that is, they would not offend, irritate, or embarrass a respondent. As in other forms of testing, items that yield variability of response or *response range* are the most useful. D. P. Campbell and J. C. Hansen (1981), for example, indicate that items such as "funeral director" and "geography" were eliminated because almost everyone indicates "dislike" to the former and "like" to the latter. An item such "college professor" on the other hand yields "like" responses of about 5% in samples of farmers to 99% in samples of behavioral scientists.

Other criteria were also used in judging whether an item would be retained or eliminated. Both predictive and concurrent validity are important and items showing these aspects were retained. For example, the Strong should have content validity and so the items should cover a wide range of occupational content. Because sex-role bias was of particular concern, items were modified (policeman became police officer) or otherwise changed. Items that showed a significant gender difference in response were not necessarily eliminated, as the task is to understand such differences rather than to ignore them. Because the United States is such a conglomeration of minorities, and because the Strong might be useful in other cultures, items were retained if they were not "culture bound," although the actual operational definition of this criterion might be a bit difficult to give. Other criteria, such as reading level, lack of ambiguity, and current terminology, were also used.

Scale development. Let's assume you want to develop an occupational scale for "golf instructors." How might you go about this? J. C. Hansen (1986) indicates that there are five steps in the construction of an occupational scale for the Strong:

1. You need to collect an occupational sample, in this case, golf instructors. Perhaps you might identify potential respondents through some major sports organization, labor union, or other societies that might provide such a roster. Your potential respondents must however, satisfy several criteria (in addition to filling out the Strong): they must be satisfied with their occupation, be between the ages of 25 and 60, have at least 3 years of experience in that occupation, and perform work that is "typical" of that occupation – for example, a golf instructor who spends his or her time primarily designing golf courses would be eliminated.
2. You also need a reference group – although ordinarily you would use the available data based on 300 "men in general" and 300 "women in general." This sample has an average age of 38 years, represents a wide variety of occupations, half professional and half nonprofessional.
3. Once you've collected your data, you'll need to compare for each of the 325 Strong items, the percent of "like," "indifferent," or "dislike" responses. The aim here is to identify 60 to 70 items that show a response difference of 16% or greater.
4. Now you can assign scoring weights to each of the 60 to 70 items. If the golf instructors endorsed "like" more often than the general sample, that item is scored +1; if the golf instructors endorsed "dislike" more often, then the item is scored −1 (for like). If there are substantial differences between the two samples on the "indifferent" response, then that response is also scored.
5. Now you can obtain the raw scores for each of your golf instructors, and compute your normative data, changing the raw scores to *T* scores.

Development. In more general terms then, the occupational scales on the Strong were developed by administering the Strong pool of items to men and women in a specific occupation and comparing the responses of this *criterion* group with those of men, or women, in general. Although the various criterion groups were different depending on the occupation, they were typically large, with Ns over 200, and more typically near 400. They were composed of individuals between the ages of 25 and 55, still active in their occupation, who had been in that occupation for at least 3 years and thus presumably satisfied, who indicated that they liked their work, and who had met some minimum level of proficiency, such as licensing, to eliminate those who might be incompetent.

The *comparison* group, the men-in-general or women-in-general sample is a bit more difficult to define, because its nature and composition has changed over the years. When Strong began his work in the mid-1920s, the in-general sample consisted of several thousand men he had tested. Later he collected a new sample based on U.S. Census Bureau statistics, but the sample contained too many unskilled and semiskilled men. When response comparisons of a criterion group were made to this comparison group, the result was that professional men shared similar interests among themselves as compared with nonprofessional men. The end result would have been a number of overlapping scales that would be highly intercorrelated and therefore of little use for career guidance. For example, a physician scale would have reflected the differences in interests between men in a professional occupation and men in nonprofessional occupations; a dentist scale would have reflected those same differences.

From 1938 to 1966 the in-general sample was a modification of the U.S. Census Bureau sample, but included only those men whose salary would have placed them in the middle class or above. From 1966 onward, a number of approaches were used, including a women-in-general sample, composed of 20 women in each of 50 occupations, and men-in-general samples with occupation membership weighted equally, i.e., equal number of biologists, physicians, life insurance salesmen, etc.

Administration. The Strong is not timed and takes about 20 to 30 minutes to complete. It can be administered individually or in groups, and is basically a self-administered inventory. The separate answer sheet must be returned to the publisher for computer scoring.

Scoring. The current version of the Strong needs to be computer scored and several such services are available. The Strong yields five sets of scores:

1. Administrative Indices
2. General Occupational Themes
3. Basic Interest Scales
4. Occupational Scales
5. Special Scales

The Administrative Indices are routine clerical checks performed by the computer as the answer sheet is scored; they are designed to assess procedural errors and are for use by the test administrator to determine whether the test results are meaningful. These indices include the number of items that were answered, the number of infrequent responses given, and the percentages of like, dislike, and indifferent responses given for each of the sections. For example, one administrative index is simply the total number of responses given. There are 325 items, and a respondent may omit some items, or may unintentionally skip a section, or may make some marks that are too light to be scored. A score of 310 or less alerts the administrator that the resulting profile may not be valid.

The General Occupational Themes are a set of six scales each designed to portray a "general" type as described in Holland's theory (discussed next). These scales were developed by selecting 20 items to represent each of the 6 types. The items were selected on the basis of both face and content validity (they covered the typological descriptions given by Holland); and statistical criteria such as item-scale correlations.

The Basic Interest Scales consist of 23 scales that cover somewhat more specific occupational areas such as, agriculture, mechanical activities, medical service, art, athletics, sales, and office practices. These scales were developed by placing together items that correlated .30 or higher with each other. Thus these scales are homogeneous and very consistent in content.

The 211 Occupational Scales in the 1994 revision cover 109 different occupations, from accountants to YMCA Directors, each scale developed empirically by comparing the responses of men and/or women employed in that occupation with the responses of a reference group of men, or of women, in general. For most of the occupations there is a scale normed on a male sample and a separate scale normed on a female sample. Why have separate gender scales? The issue of gender differences is a complex one, fraught with all sorts of social and political repercussions. In fact, however, men and women respond differently to about half of the items contained in the Strong, and therefore separate scales and separate norms are needed (J. C. Hansen & D. P. Campbell, 1985). Most of these samples were quite sizable, with an average close to 250 persons, and a mean age close to 40 years; to develop and norm these scales more than 142,000 individuals were tested. Some of the smaller samples are quite unique and include astronauts, Pulitzer Prize-winning authors, college football coaches, state governors, and even Nobel prize winners (D. P. Campbell, 1971). Because these scales have been developed empirically, they are factorially complex, most made up of rather heterogeneous items, and often with items that do not have face validity. The Psychologist Scale, for example, includes items that reflect an interest in science, in the arts, and in social service, as well as items having to do with business and military activities, which are weighted negatively. Thus two people with identical scores on this scale, may in fact have different patterns of responding. Though empirically these scales work well, it is difficult for a counselor to understand the client unless one undertakes an analysis of such differential responding. However, by looking at the scores on the Basic Interest Scales, mentioned above, one can better determine where the client's interests lie, and thus better understand the results of the specific occupational scales.

Finally, there are the Special Scales. At present, two of these are included in routine scoring:

1. The Academic Comfort Scale which was developed by contrasting the responses of high-GPA students with low-GPA students; this scale attempts to differentiate between people who enjoy being in an academic setting and those who do not.
2. The Introversion-Extroversion scale that was developed by contrasting the responses of introverted with those of extroverted individuals, as defined by their scores on the MMPI scale of the same name. High scorers (introverts) prefer working with things or ideas, while low scorers (extroverts) prefer working with people.

Scores on the Strong are for the most part presented as *T* scores with a mean of 50 and SD of 10.

Interpretation of the profile. The resulting Strong profile presents a wealth of data, which is both a positive feature and a negative one. The negative aspect comes about because the wealth of information provides data not just on the career interests of the client, but also on varied aspects of their personality, their psychological functioning, and general psychic adjustment, and thus demands a high degree of psychometric and psychological sophistication from the counselor in interpreting and communicating the results to the client. Not all counselors have such a degree of training and sensitivity, and often the feedback session to the client is less than satisfying (for some excellent suggestions regarding test interpretation and some illustrative case studies, see D. P. Campbell & J. C. Hansen, 1981).

Criterion-keying. When the Strong was first introduced in 1927, it pioneered the use of *criterion-keying* of items, later incorporated into personality inventories such as the MMPI and the CPI. Thus the Strong was administered to groups of individuals in specific occupations, and their responses compared with those of "people in general." Test items that showed differential response patterns between a particular occupational group, for example dentists, and people in general then became the dentist scale. Hundreds of such occupational scales were developed, based on the simple fact that individuals in different occupations have different career interests. It is thus possible to administer the Strong to an individual and determine that person's degree of similarity between their career interests and those shown by individuals in specific careers. Thus each of the occupational scales is basically a subset of items that show large differences in response percentages between individuals in that occupation and a general sample. How large is large? In general, items that show at least a 16%

difference are useful items; for example, if 58% of the specific occupational sample respond "like" to a particular item vs. 42% of the general sample, that item is potentially useful (D. P. Campbell & J. C. Hansen, 1981). Note that one such item would not be very useful, but the average occupational sample scale contains about 60 such items, each contributing to the total scale.

Gender bias. The earlier versions of the Strong not only contained separate scoring for occupations based on the respondent's gender, but the separate gender booklets were printed in blue for males and pink for females! Thus, women's career interests were compared with those of nurses, school teachers, secretaries and other traditionally "feminine" occupations. Fortunately, current versions of the Strong have done away with such sexism, have in fact pioneered gender equality in various aspects of the test, and provide substantial career information for both genders, and one test booklet.

Holland's theory. The earlier versions of the Strong were guided primarily by empirical considerations, and occupational scales were developed because there was a need for such scales. As these scales proliferated, it became apparent that some organizing framework was needed to group subsets of scales together. Strong and others developed a number of such classifying schemas based on the intercorrelations of the occupational scales, on factor analysis, and on the identification of homogeneous clusters of items. In 1974, however, a number of changes were made, including the incorporation of Holland's (1966; 1973; 1985a) theoretical framework as a way of organizing the test results.

Holland believes that individuals find specific careers attractive because of their personalities and background variables; he postulated that all occupations could be conceptualized as representing one of six general occupational themes labeled realistic, investigative, artistic, social, enterprising, and conventional.

Individuals whose career interests are high in the *realistic* area are typically aggressive persons who prefer concrete activities to abstract work. They prefer occupations that involve working outdoors and working with tools and objects rather than with ideas or people. These individuals are typically practical and physically oriented but may have difficulties expressing their feelings and concerns. They are less sociable and less given to interpersonal interactions. Such occupations as engineer, vocational agriculture teacher, and military officer are representative of this theme.

Individuals whose career interests are high in the *investigative* theme focus on science and scientific activities. They enjoy investigative challenges, particularly those that involve abstract problems and the physical world. They do not like situations that are highly structured, and may be quite original and creative in their ideas. They are typically intellectual, analytical, and often quite independent. Occupations such as biologist, mathematician, college professor, and psychologist are representative of this theme.

As the name implies, the *artistic* theme centers on artistic activities. Individuals with career interests in this area value aesthetics and prefer self-expression through painting, words, and other artistic media. These individuals see themselves as imaginative and original, expressive, and independent. Examples of specific careers that illustrate this theme are artist, musician, lawyer, and librarian.

The fourth area is the *social* area; individuals whose career interests fall under this theme are people-oriented. They are typically sociable and concerned about others. Their typical approach to problem solving is through interpersonal processes. Representative occupations here are guidance counselor, elementary school teacher, nurse, and minister.

The *enterprising* area is the area of sales. Individuals whose career interests are high here see themselves as confident and dominant, like to be in charge, and to persuade others. They make use of good verbal skills, are extroverted, adventurous, and prefer leadership roles. Typical occupations include store manager, purchasing agent, and personnel director.

Finally, the *conventional* theme focuses on the business world, especially those activities that characterize office work. Individuals whose career interests are high here are said to fit well in large organizations and to be comfortable working within a well-established chain of command, even though they do not seek leadership positions. Typically, they are practical and sociable, well controlled and conservative. Representative occupations are those of accountant, secretary, computer operator, and credit manager.

As the description of these types indicates, Holland's model began its theoretical life as a personality model. Like other personality typologies that have been developed, it is understood that "pure" types are rare. But the different types are differentiated: A person who represents the "conventional" type is quite different from the person who is an "artistic" type.

Finally, there is a congruence between personality and occupation resulting in satisfaction. An artistic type of person will most likely not find substantial satisfaction in being an accountant. Holland's theory is not the only theory of career development, but has been one of the most influential, especially in terms of psychological testing (for other points of view see Bergland, 1974; Gelatt, 1967; Krumboltz, Mitchell, & Gelatt, 1975; Osipow, 1983; Tiedeman & O'Hara, 1963).

Reliability. The reliabilities associated with the Strong are quite substantial. D. P. Campbell and J. C. Hansen (1981), for example, cite median test-retest correlations, with a 2-week interval the $r = .91$, with a 2- to 5-year interval, the *r*s range from .70 to .78, and with a 20+ year interval, the *r*s range from .64 to .72. Not only is the Strong relatively stable over time, so are career interests.

Test-retest reliabilities for the Basic Interest Scales are quite substantial, with median coefficients of .91 for a 2-week period, .88 for 1 month-, and .82 for 3-year periods. Test-retest correlations also vary with the age of the sample, with the results showing less reliability with younger samples, for example 16-year-olds, as might be expected.

Validity. The Basic Interest Scales have substantial content and concurrent validity; that is, their content makes sense, and a number of studies have shown that these scales do indeed discriminate between persons in different occupations. In general, their predictive validity is not as high, and some scales seem to be related to other variables rather than occupational choice; for example, the Adventure Scale seems to reflect age, with older individuals scoring lower.

Strong was highly empirically oriented and developed not just an inventory, but a rich source of longitudinal data. For example, after the SVIB was published, he administered the inventory to the senior class at Stanford University, and 5 years later contacted them to determine which occupations they had entered, and how these occupations related to their scores on the inventory.

The criterion then for studying the predictive validity of the Strong becomes the occupation that the person eventually enters. If someone becomes a physician and their Strong profile indicates a high score on the Physician scale, we then have a "hit." The problem, however, is that the world is complex and individuals do not necessarily end up in the occupation for which they are best suited, or which they desire. As Strong (1935) argued, if final occupational choice is an imperfect criterion, then a test that is validated against such a criterion must also be imperfect. This of course is precisely the problem we discussed in Chapter 3; a test cannot be more valid than the criterion against which it is matched, and in the real world there are few, if any, such criteria. Nevertheless, a number of studies both by Strong (1955) and others (e.g., D. P. Campbell, 1971; Dolliver, Irwin, & Bigley, 1972) show substantial predictive validity for the Strong, with a typical hit rate (agreement between high score on an Occupational Scale and entrance into that occupation) of at least 50% for both men and women. There is of course something reassuring that the hit rates are not higher; for one thing it means that specific occupations do attract people with different ideas and interests, and such variability keeps occupations vibrant and growing.

Faking. In most situations where the Strong is administered, there is little if any motivation to fake the results because the client is usually taking the inventory for their own enhancement. There may be occasions, however, when the Strong is administered as part of an application process; there may be potential for faking in the application for a specific occupation or perhaps entrance into a professional school.

Over the years, a number of investigators have looked at this topic, primarily by administering the Strong twice to a sample of subjects, first under standard instructions, and secondly with instructions to fake in a specific way, for example, "fake good to get higher scores on engineering" (e.g., Garry, 1953; Wallace, 1950). The results basically support the notion that under such instructions Strong results can be changed. Most of these studies represent artificial situations where captive subjects are instructed to fake. What happens in real life? D. P. Campbell (1971) reports the results of a doctoral dissertation that compared the Strong profiles of 278 University of Minnesota males who had completed the Strong first for counseling purposes and later had completed the Strong a second time as part of their application procedure to the University of Minnesota medical school. Presumably, when the Strong was taken for counseling purposes the respondents completed the inventory honestly, but when the Strong was taken as part of an application process, faking might have occurred, especially on those items possibly related to a career in medicine. In fact, for 47% of the sample, there was no difference on their physician scale score between the two administrations. For 29%, there was an increase, but not substantial. For 24%, there was a substantial increase, enough to have a "serious effect" on its interpretation by an admissions officer. Of course, just because there was an increase does not mean that the individual faked; the increase might well reflect legitimate growth in medical interest. There are three points to be made here: (1) faking is possible on the Strong, (2) massive distortions do not usually occur, (3) the resulting profile typically shows considerable consistency over time.

Inconsistencies. Because the Strong contains different sets of scales developed in different ways, it is not unusual for a client's results to reflect some inconsistencies. R. W. Johnson (1972) reported that some 20% of profiles have at least one or more such inconsistencies between Occupational Scales and Basic Interest Scales. D. P. Campbell and J. C. Hansen (1981) argue that such inconsistencies are meaningful and result in more accurate test interpretation because they force both the counselor and the client to understand the meaning of the scales and to go beyond the mere occupational label. For example, the Basic Interest Scales reflect not only career interests but leisure interests as well (Cairo, 1979).

Unit weighting. The Strong illustrates nicely the concept of unit weights as opposed to variable weights. Let's suppose we are developing a scale for a new occupation of "virtual reality trainer" (VRT). We administer the Strong, which represents a pool of items and an "open" system, to a group of VRTs and a group of "people in general," and identify those items that statistically separate the two groups.

Let's say for example, that 85% of our VRTs indicate like to the item "computer programmer" vs. only 10% for the general sample, and that 80% of the VRTs also indicate dislike to the item "philosopher" vs. 55% for the general sample. Both items show a significant difference in response pattern and so both would be included in our scale. But clearly, one item is more "powerful," one item shows a greater difference between our two groups, and so we might logically argue that such an item should be given greater weight in the way the scale is scored. That indeed is what Strong originally did; the items were weighted based on a ratio of the response percentage of the specific occupational sample vs. the response percentage of the general sample. And so initially, Strong items were scored with weights ranging from -30 to $+30$. Such scoring, especially in the precomputer days, was extremely cumbersome, and so was simplified several times until in 1966 the weights of $+1$, 0, or -1 were used. Empirical studies of unit weights vs. variable weights show the unitary weights to be just as valid.

Percentage overlap. Another interesting concept illustrated by the Strong is that of *percentage overlap*. Let's assume we have administered the Strong to two groups of individuals, and we are interested in looking at a specific occupational scale for which our theory dictates the two samples should differ. How do we determine whether the two groups differ? Ordinarily we would carry out a *t* test or an analysis of variance to assess whether the means of the two groups are statistically different from each other (you recall, by the way, that when we have two groups, the two procedures are the same in that $t^2 = F$). Such a procedure tells us that yes (or no) there is a difference, but it doesn't really tell us how big that difference is, and does not address the issue of practicality – a small mean difference could be statistically significant if we have large enough samples, but would not necessarily be useful.

A somewhat different approach was suggested by Tilton (1937) who presented the statistic of

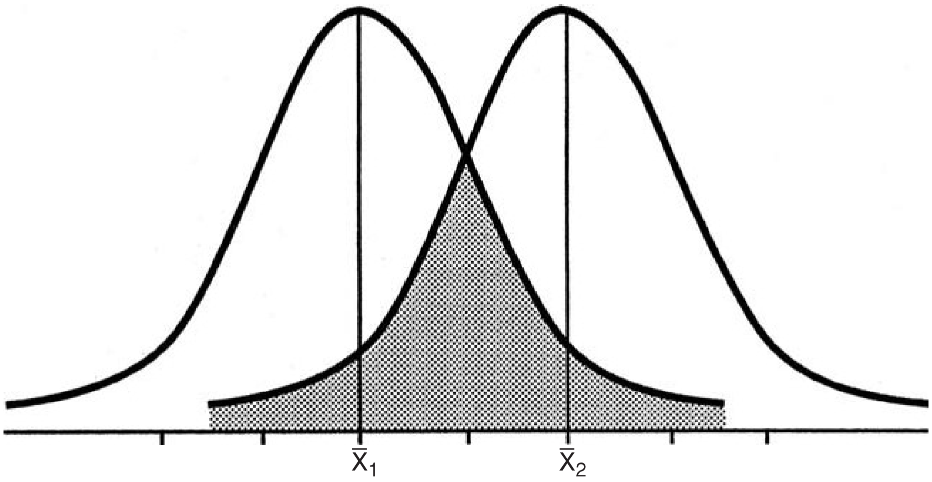


FIGURE 6–3. Two distributions separated from each other by two standard deviations.

percent overlap, which is simply the percentage of scores in one sample that are matched by scores in the second sample. If the two distributions of scores are totally different and therefore don't overlap, the statistic is zero. If the two distributions are identical and completely overlap, then the statistic is 100%. If the intent of a scale is to distinguish between two groups, then clearly the lower the percentage overlap, the more efficient (valid) is the scale. Tilton called this statistic the *Q* index, and it is calculated as follows:

$$Q = \frac{M_1 - M_2}{2(\mathrm{SD}_1 + \mathrm{SD}_2)}$$

Once *Q* is computed, the percent overlap can be determined using Tilton's (1937) table. Essentially, the *Q* index is a measure of the number of standard deviation units that separate the two distributions. For example, a *Q* value of 2 represents two distributions that are separated from each other by two standard deviations, and have an overlap of about 32%. Figure 6.3 illustrates this. Note that if our occupational scale were an IQ test, the means of the two groups would differ by about 30 points – a rather substantial difference. The median percent overlap for the Strong occupational scales is in fact about 34%. This is of course a way of expressing concurrent validity. Scales that reflect well-defined occupations such as physicist or chemist, have the lowest overlap or highest validity. Scales that assess less well-defined occupations, such as that of college professor, have a higher degree of overlap and, therefore, lower concurrent validity.

Racial differences. Although racial differences on the Strong have not been studied extensively, and in fact the SVIB Handbook (D. P. Campbell, 1971) does not discuss this topic, the available studies (e.g., Barnette & McCall, 1964; Borgen & Harper, 1973) indicate that the Strong is not racially biased and that its predictive validity and other psychometric aspects for minority groups are equivalent to those for whites.

Item response distribution. D. P. Campbell and J. C. Hansen (1981) indicate that interest measurement is based on two empirical findings: (1) different people give different responses to the individual items; and (2) people who are satisfied with their particular occupation tend to respond to particular items in a characteristic way. Given these two statements, the *item response distribution* for a particular item charts the value of that item and its potential usefulness in the inventory. At the Center for Interest Measurement Research of the University of Minnesota extensive data on the Strong is stored on computer archives, going back to the original samples tested by Strong. For example, D. P. Campbell and J. C. Hansen (1981) show the item response distribution for the item "artist" given by some 438 samples, each sample typically ranging from less than 100 to more than 1,000 individuals in a specific occupation. Both male and female artist samples tend to show near unanimity in their endorsement of "like"; at the other extreme, male farmers show an 11% "like" response, and females in life insurance sales a 32% "like" response.

Longitudinal studies. The Strong has been used in a number of longitudinal studies, and specifically to assess the stability of vocational interests within occupations over long time spans. D. P. Campbell (1966) asked and answered three basic questions: (1) Do Strong scales developed in the 1930s hold up in cross-validation years later? The answer is yes; (2) When Strong scales have been revised, did the revised scales differ drastically from the originals? The answer is not much; (3) Do the individuals of today who hold the same job as the individuals in Strong's criterion groups of the 1930s have the same interest patterns? The answer is pretty much so.

Inventoried vs. expressed interests. "Inventoried" interests are assessed by an inventory such as the Strong. "Expressed" interests refer to the client's direct comments such as, "I want to be an engineer" or "I am going to study environmental law." How do these two methods compare? If for example, the Strong were simply to mirror the client's expressed interests, why waste time and money, when the same information could be obtained more directly by simply asking the subject what they want to be. Of course, there are many people who do not know what career to pursue, and so one benefit of the Strong and similar instruments, is that it provides substantial exploratory information. Berdie (1950) reported correlations of about .50 in studies that compared inventoried and expressed interests. However, Dolliver (1969) pointed out that this deceptively simple question actually involves some complex issues including the reliability and validity of both the inventory and the method by which expressed interests are assessed, attrition of subjects upon follow-up, and the role of chance in assessing such results.

The Kuder Inventories

Introduction. A second set of career-interest inventories that have dominated psychological testing in this area, has been the inventories developed by Frederic Kuder. There are actually three Kuder inventories: (1) the Kuder Vocational Preference Record (KVPR), which is used for career counseling of high-school students and adults; (2) the Kuder General Interest Survey (KGIS), which is a downward extension of the KVPR, for use with junior and senior high-school students in grades 6 through 12; and (3) the Kuder Occupational Interest Survey (KOIS), designed for grades 10 through adulthood. The first two yield scores in 10 general areas, namely: artistic, clerical, computational, literary, mechanical, musical, outdoor, persuasive, scientific, and social service. The third, the KOIS, yields substantially more information, and our discussion will focus primarily on this instrument. Once again, we use the term "Kuder" as a more generic designation, except where this would violate the meaning.

Development. Initially, the Strong and the Kuder represented very different approaches. The Strong reflected criterion-group scaling while the Kuder represented homogeneous scaling, that is clustering of items that are related. Over the years however, the two approaches have borrowed heavily from each other, and thus have become more convergent in approach and process.

Description. The KOIS takes about 30 minutes to complete and is not timed. It can be administered to one individual or to a large group at one sitting. Like the Strong, it too must be computer scored. The KOIS is applicable to high-school students in the 10th grade or beyond (Zytowski, 1981). In addition to 126 occupational scales, the KOIS also has 48 college-major scales. The KOIS also has a number of validity indices, similar to the Strong's administrative indices, including an index that reflects number of items left blank, and a *verification* score that is basically a "fake good" scale. As with most other major commercially published tests, there is not only a manual (Kuder & Diamond, 1979), but additional materials available for the practitioner (e.g., Zytowski, 1981; 1985).

Scale development. We saw that in the Strong, occupational scales were developed by pooling those 40 to 60 items in which the response proportions of an occupational group and an in-general group differed, usually by at least 16%. The Kuder took a different approach. The Kuder was originally developed by administering a list of statements to a group of college students, and based on their responses, placing the items into 10 homogeneous scales. Items within a scale correlated highly with each other, but not with items in the other scales. Items were then placed in triads, each triad reflecting three different scales. The respondent indicates which item is most preferred and which item is least preferred. Note that this results in an ipsative instrument – one cannot obtain all high scores. To the degree that one scale score is high, the other scale scores must be lower.

Let's assume we are developing a new occupational scale on the Kuder for "limousine driver," and find that our sample of limousine drivers endorses the first triad as follows:

Item #	Most preferred	Least preferred
1	20%	70%
2	60%	15%
3	20%	15%

That is, 20% of our sample selected item #1 as most preferred, 60% selected item 2, and 20% item 3; similarly, 70% selected item #1 as least preferred, 15% item 2, and 15% item 3.

If you were to take the Kuder, your score for the first triad on the "limousine driver" scale would be the proportion of the criterion group that endorsed the same responses. So if you indicated that item #1 is your most preferred and item #2 is your least preferred, your score on that triad would be .20 + .15 = .35. The highest score would be obtained if you endorsed item #2 as most and item #1 as least; your score in this case would be .60 + .70 = 1.30. Note that this triad would be scored differently for different scales because the proportions of endorsement would presumably change for different occupational groups. Note also that with this approach there is no need to have a general group.

In any one occupational group, we would expect a response pattern that reflects homogeneity of interest, as in our fictitious example, where the majority of limousine drivers agree on what they prefer most and prefer least. If we did not have such unanimity we would expect a "random" response pattern, where each item in the triad is endorsed by approximately one-third of the respondents. In fact, we can calculate a total score across all triads that reflects the homogeneity of interest for a particular group, and whether a particular Kuder scale differentiates one occupational group from others (see Zytowski & Kuder, 1986, on how this is done).

Scoring. Scales on the KOIS are scored by means of a "lambda" score, which is a modified biserial correlation coefficient and is essentially an index of similarity between a person's responses and the criterion group for each scale. Rather than interpreting these lambda scores directly, they are used to rank order the scales to show the magnitude of similarity. Thus, the profile sheet that summarizes the test results is essentially a listing of general occupational interests (e.g., scientific, artistic, computational), and of occupations and of college majors, all listed in decreasing order of similarity.

Reliability. Test-retest reliabilities seem quite acceptable, with for example median reliability coefficients in the .80s over both a 2-week and a 3-year period (Zytowski, 1985).

Validity. Predictive validity also seems to be acceptable, with a number of studies showing about a 50% congruence between test results and subsequent entrance into an occupation some 12 to 19 years later (Zytowski & Laing, 1978).

Other Interest Inventories

A large number of other inventories have been developed over the years, although none have reached the status of the Strong or the Kuder. Among the more popular ones are the Holland Self-Directed Search (Holland, 1985b), the Jackson Vocational Interest Survey (D. N. Jackson, 1977), the Career Assessment Inventory (Johansson, 1975), the Unisex edition of the ACT Interest Inventory (Lamb & Prediger, 1981), and the Vocational Interest Inventory (P. W. Lunneborg, 1979).

Interest inventories for disadvantaged. A number of interest inventories have been developed for use with clients who, for a variety of reasons, may not be able to understand and/or respond appropriately to verbal items such as those used in the Strong and the Kuder. These inventories, such as the Wide Range Interest Opinion Test (J. F. Jastak & S. R. Jastak, 1972) and the Geist Picture Interest Inventory (Geist, 1959) use drawings of people or activities related to occupational tasks such as doing laundry, taking care of animals, serving food, and similar activities. Some of these inventories use a forced-choice format, while others ask the respondent to indicate how much they like each activity. Most of these have adequate reliability but leave much to be desired in the area of validity.

Interest inventories for nonprofessional occupations. Both the Strong and the Kuder have found their primary application with college-bound students and adults whose expectations are to enter professions. In fact, most career-interest inventories are designed for occupations that are entered by middle-class individuals. In large part, this reflects a reality of our culture, that perhaps is changing. At least in the past, individuals from lower socioeconomic classes did not have much choice, and job selection was often a matter of availability and financial need. For upper socioeconomic class individuals, their choice was similarly limited by family expectations and traditions, such as continuing a family business or family involvement in government service. A number of other interest inventories have been developed that are geared more for individuals entering nonprofessional occupations.

One example of such inventories is the Career Assessment Inventory (CAI; Johansson, 1986), first introduced in 1975 and subsequently revised several times, recently to include both nonprofessional and professional occupations. The CAI currently contains some 370 items similar in content to those of the Strong and takes about 40 to 45 minutes to complete. For each item, the client responds on a 5-point scale ranging from "like very much" to "dislike very much." Like the Strong, the CAI contains the six general-theme scales that reflect Holland's typology, 25 Basic Interest scales (e.g., electronics, food service, athletics-sports), and 111 occupational scales (such as accountant, barber/hairstylist, carpenter, fire-fighter, interior designer, medical assistant, police officer, and truck driver). Although the CAI seems promising, in that it was well-designed psychometrically and shows adequate reliability, it too has been criticized, primarily for lack of evidence of validity (McCabe, 1985; Rounds, 1989).

Other career aspects. In addition to career interests, there are a number of questionnaires designed to assess a person's attitudes, competencies, and decision-making skills, all in relation to career choice. For example, the Career Decision Scale (Osipow, 1987) is an 18-item scale designed to assess career indecision in college students, and the Career Maturity Inventory (Crites, 1978) is designed to assess career-choice competencies (such as self-knowledge and awareness of one's interests) and attitudes (such as degree of independence and involvement in making career decisions).

Lack of theory. One criticism of the entire field of career-interest measurement is that it has been dominated by an empirical approach. The approach has been highly successful, yet it has resulted in a severe lack of theoretical knowledge about various aspects of career interests. For example, how do these interests develop? What psychological processes mediate and affect such interests? How do such variables as personality, temperament, and motivation relate to career interests? There is now a need to focus on construct validity rather than criterion validity. To be sure, such questions have not been totally disregarded. For example, Roe (Roe & Klos, 1969; Roe & Siegelman, 1964) felt that career choices reflected early upbringing and that children who were raised in an accepting and warm family atmosphere would choose people-oriented occupations. Others have looked to a genetic component of career interests (e.g., Grotevant, Scarr, & Weinberg, 1977).

New occupational scales. The world of work is not a static one, and especially in a rapidly expanding technology, new occupations are created. Should we therefore continue to develop new occupational scales? Some authors (e.g., Borgen, 1986; Burisch, 1984) have argued that a simple, deductive approach to career interest measurement may be now more productive than the empirical and technical development of new scales. These authors believe that we now have both the theories and the empirical knowledge related to the occupational world, and we should be able to locate any new occupation in that framework without needing to go out and develop a new scale. Indeed, Borgen (1986) argues that occupational scales may not be needed and that a broad perspective, such as the one provided by Holland's theory, is all that is needed.

SUMMARY

We have looked at the measurement of attitudes, values, and interests. From a psychometric point of view these three areas share much in common, and what has been covered under one topic could, in many instances be covered under a different topic. We looked at four classical methods to construct attitude scales; the method of equal appearing intervals or Thurstone method, the method of summated ratings or Likert method, the Bogardus social distance scale, and Guttman scaling. In addition, we looked at the Semantic Differential, checklists, numerical and graphic rating scales, and self-anchoring scales.

In the area of values, we looked at the Study of Values, a measure that enjoyed a great deal of popularity years ago, and the Rokeach Value Survey, which is quite popular now. We also briefly discussed the Survey of Interpersonal Values and the Survey of Personal Values to illustrate another approach. In the area of career interests, we focused primarily on the Strong and the Kuder inventories that originally represented quite different approaches but in recent revisions have become more alike.

SUGGESTED READINGS

Campbell, D. P. (1971). An informal history of the SVIB. In *Handbook for the Strong Vocational Interest Blank* (pp. 343–365). Stanford, CA: Stanford University Press.

This is a fascinating account of the SVIB, from its early beginnings in the 1920s to the mid 1960s, shortly after Strong's death. For those who assume that computers have been available "forever," this chapter has a wonderful description of the challenges required to "machine score" a test.

Domino, G., Gibson, L., Poling, S., & Westlake, L. (1980). Students' attitudes towards suicide. *Social Psychiatry*, *15*, 127–130.

The investigators looked at the attitudes that college students have toward suicide. They used the Suicide Opinion Questionnaire, and administered it to some 800 college students in nine different institutions. An interesting study illustrating the practical application of an attitude scale.

Kilpatrick, F. P. & Cantril, H. (1960). Self-anchoring scaling: A measure of individual's unique reality worlds. *Journal of Individual Psychology*, *16*, 158–173.

An interesting report where the two authors present the self-anchoring methodology and the results of several studies where such scales were administered to adult Americans, legislators from seven different countries, college students in India, and members of the Bantu tribe in South Africa.

Lawton, M. P. & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist*, *9*, 179–186.

Two scales are presented for use with institutionalized elderly. The first scale focuses on physical self-maintenance and covers six areas, toilet use, feeding, dressing, grooming, and physical ambulation. The second scale, Instrumental Activities of Daily Living, covers eight areas ranging from the ability to use the telephone to the ability to handle finances. The scale items are given in this article and clearly illustrate the nature of Guttman scaling, although the focus is on how the scales can be used in various settings, rather than in how the scales were developed.

Rokeach, M. & Ball-Rokeach, S. J. (1989). Stability and change in American value priorities, 1968–1981. *American Psychologist*, *44*, 775–784.

Psychological tests can be useful not only to study the functioning of individuals, but to assess an entire society. In this report, the authors analyze national data on the RVS which was administered by the National Opinion Research Center of the University of Chicago in 1968 and again in 1971, and by the Institute for Social Research at the University of Michigan in 1974 and in 1981. Although there seems to be remarkable stability of values over time, there were also some significant changes – for example, "equality" decreased significantly.

DISCUSSION QUESTIONS

1. What are some of the strengths and weaknesses of attitude scales?
2. Which of the various ways of assessing reliability would be most appropriate for a Guttman scale?
3. What might be some good bipolar adjectives to use in a Semantic Differential scale to rate "my best teacher"?
4. What are the basic values important to college students today? Are these included in the Rokeach?
5. Most students take some type of career-interest test in high school. What is your recollection of such a test and the results?