Measurement

Charles M. Judd **Gary H. McClelland**

University of Colorado

DRAFT: Please do not quote or cite without permisson.

To appear in Gilbert, D., Fiske, S., & Lindzey, G., Handbook of Social Psychology. Preparation of this chapter was supported by NIMH grant R01 MH45049 to Charles M. Judd and Bernadette Park. Correspondence should be addressed to either author at the Department of Psychology, University of Colorado, Boulder, CO 80309-0345 or via the Internet to charles.judd or gary.mcclelland both @colorado.edu.

Table of Contents

Introduction

> *Although measurement is one of the gods modern psychologists pay homage to with great regularity, the subject of measurement remains as elusive as ever*. (Suppes & Zinnes, 1963)

Despite considerable work on measurement, this statement by Suppes and Zinnes remains as true today as it was in 1963. The goal of this chapter is to provide practical advice for constructing and testing measures and scales in social psychology. But the reader should not expect to find a simple, step-by-step, how-to manual, because measurement does indeed remain elusive. Much of this chapter is concerned with the *how* of measurement and scaling. Before considering specific methods and procedures, however, we should consider *why* as social scientists we want to measure things and we should define *what* we mean by "measurement." The next section addresses these two questions.

The What and the Why of Measurement

Most books and chapters on measurement jump to the how-to without considering what measurement really is, why we are interested in it, and what its benefits are. There are of course some notable exceptions. Torgerson (1958), Coombs (1964), and Messick (1989), for instance, consider the role and benefits of measurement and scaling within the larger context of the goals of science. The material in this section is heavily influenced by these and other sources.

The raw data of social psychology, and of the social and behavioral sciences in general, potentially consist of infinitely minute observations of ongoing behavior and attributes of individuals, social groups, social environments, and other entities or objects that populate the social world. Measurement is the process by which these infinitely varied observations are reduced to compact descriptions or models that are presumed to represent meaningful regularities in the entities that are observed.

This definition is rather different from the classic definition of measurement offered by Stevens in 1951 (p. 22). He defined measurement simply "as the assignment of numerals to objects or events according to rules." Like Torgerson (1958) and Dawes and Smith (1985), we take issue with this definition on at least two grounds. First of all, not all measurement involves numbers. As we illustrate later on, compact descriptions of observations can be well represented non-numerically. Second, and more importantly, rules for assigning numbers constitute measurement only if the subsequent numbers end up representing something of meaning, some regularity of attributes or behaviors that permits prediction. Dawes and Smith (1985; p. 511) nicely make the point:

> *Consider, for example, the number of rules that could be used to assign numbers to beauty contestants or politicans (e.g., cube the distance between the candidate's chin and left forefinger when he or she is tanding at attention; then divide by the time of his or her birth). Few of these rules would tell us anyting worth knowing (e.g., who will win). The assignment of numbers not only must be orderly if it is to yield measurment but must also represent meaningful attributes and yield meaningful predictions.*

The compact model or description that we construct of observations through measurement we will call a scale or a variable. The meaningful attribute or regularity that it is presumed to represent we will call a construct. Accordingly, measurement consists of rules that assign scale or variable values to entities to represent the constructs that are thought to be theoretically meaningful.

The entities to be measured typically consist of individuals or social groups in social psychology. But in fact, many other sorts of objects can be measured. For example, the entities to be measured might be atttitude statements, cookies made with varying amounts of sugar and salt, archaeological sites, political candidates, or automobiles. Likewise, the constructs to be measured may include a wide array of attributes thought to be meaningful within the context of some theory about the entities. Potential constructs or attributes to be measured about these entities might be (in order of the above entity examples) political liberalness, taste appeal, age, honesty, or fuel efficiency.

A bit more abstractly, assume that *a, b, c*, ..., represent different entities and we want to measure one or more constructs or attributes of these entities. We will say that *s()* is a scale if it assigns values to all the entities under consideration, with *s(a), s(b), s(c), …* being the respective values. This scale will then constitute a measurement of the entities if in fact it adequately models or represents the construct that we wish to measure.

As an example consider the following specific rule for assigning numbers to attitude statement to represent how "liberal" they are:

> Assignment Rule. The number assigned to attitude statement a will be larger than the number assigned to attitude statement b, that is, $s(a) > s(b)$ if and only if a majority of a panel of 15 randomly-selected students judges that statement a is more liberal than statement b.

This assignment rule is a sufficient example of measurement according to Stevens (1951) definition. According to ours, it is not. Further conditions of measurement are that there is in fact an attribute "liberalness" that can be assessed and that the assigned scale values actually capture or adequately model this construct. In essence, for a variable or scale to be a valid measure, the scale values of the entities must resemble the true but unknown standings of the entities on the construct that is of theoretical interest. The variable or scale is then said to possess "construct validity" (Cronbach & Meehl, 1955).

The defining conditions of measurement necessitate that measurement and theory are inextricably linked. Measurement presupposes a theory that defines the important constructs to be measured and that provides a motivation for the rule for assigning scale values to entities. Theory in turn depends on measurement, for confidence in a set of theoretical hypotheses increases by showing empirical relationships among measured variables. And theoretical disconfirmation only occurs if the relationships posited by theory are empirically not found.

Good measurement then can lead to a disconfirmation of theoretical expectations. This can happen in two ways. The first way is the most obvious: Empirical relationships among measured variables may show inconsistencies with theoretical expectations. This then leads to theoretical modifications or elaborations and further empirical work. The second way in which theoretical disconfirmation can happen is that the measurement model itself may fail. For instance, the theory may lead one to believe that there is in fact a construct to be measured or scaled, but the scaling procedure itself may reveal inconsistencies that suggest that a unidimensional construct is untenable. A good scaling or measurement model thus is one that provides the possibility of theoretical disconfirmation. It allows that measurement inconsistencies suggest that constructs do not exist as theoretically defined.

Two different measurement traditions exist within psychology as a whole and these have influenced contemporary approaches to measurement in social psychology. As we shall see, the two traditions differ in the sort of evidence that they rely upon in order to disconfirm a measurement model. We follow the lead of Dawes and Smith (1985) and others in refering to these two different traditions as the axiomatic or representational approach and the psychometric approach. In the sections that follow we briefly define these two approaches. The subsequent organization of the chapter is based on the distinction between the two approaches. Thus, following this introduction, we first discuss in some detail axiomatic measurement and its approach to model disconfirmation. We then turn to the psychometric tradition, outlining how it assesses measurement adequacy. In writing this chapter, it was striking for us to note the extent to which these two traditions speak different languages and have, as a result, failed to communicate with each other. Accordingly, in the concluding section of the chapter we attempt to build bridges between them.

Axiomatic or Representational Measurement

Axiomatic or representational measurement is the assignment of numbers to entities so that properties of the numbers (e.g., greater than, addition, subtraction, multiplication) *represent*t empirical relationships. The rule presented above for assigning scale values to attitude statements provides an example: the greater-than property of numbers represents or corresponds to the empirical observation that a majority of the panel judged one statement to be more liberal than another. Or, as another example, consider the physical measurement of length. If rod a is 5 cm and rod b is 15 cm, then placing the two rods end to end will produce a rod 15 cm long. In this case, the additive property of numbers represents the physical concatenation of the two rods. In effect, numbers along with their specified properties become a model for our observations.

An important feature of numerical representations is that they can be used to make predictions about the empirical observations. For example, if statement c has a scale value of 9 and statement d a scale value of 2, we would predict that a majority of the panel would judge statement c to be more liberal than statement d. Similarly, we can predict that concatenating rods of lengths 3 and 5 would produce exactly the same length as concatenating rods of lengths 2 and 6 because 3 + 5 = 2 + 6. Note that we must be careful in each instance to specify the properties of numbers that are supposed to apply to the empirical observations. For example, our assignment rule for attitude statements only involved the order property of numbers. We would have no basis, given this particular assignment rule, for predicting that someone who agreed with two attitude statements with scale values of 3 and 4 was more liberal than someone who agreed with a statement with a scale value of 6 even though 3 + 4 > 6. The assignment rule simply does not imply that addition is meant to represent such empirical observations.

It is usually not obvious which numerical properties ought to apply to a given set of observations. At this point, you may not even understand why the additive property should not apply in the example of the attitude statements. That is why we need theories of measurement to tell us the conditions or relationships that must be true for our observations before we can represent them with numbers and specified numerical properties. Axiomatic measurement theory is concerned with the specification and testing of those conditions for a variety of data types (Coombs, 1964). And we detail these later in the chapter.

For now, note that the ability to make predictions about the set of entities whose properties we are measuring provides us with an internal consistency check of the particular scaling model we are using. For example, if statement a is judged more liberal than statement b (which we will denote by $a \succ b$ for a dominates b) and $b \succ c$, then clearly the number assigned to a ought to be higher than the number assigned to c. After assigning numbers to a, b, and c we could then predict an observation that has not yet been used in the measurement process: namely, a should be judged to be more liberal than c. If it is, then we have a bit more confidence in our measurement and scaling; if it is not, then we have a serious problem with our measurement model because an internal consistency check has been violated. The implications of violations are discussed more thoroughly in later sections. The important point here is that a defining characteristic of representational measurement is that it is always possible to derive internal consistency checks to assess the validity of the measurement model being used to represent the observations.

Psychometric Measurement

The other approach to measurement relies on external and aggregate patterns of data to evaluate the adequacy of a measurement model rather than on internal consistency checks. The most common example of the non-representational approach to measurement in the social sciences is the ubiquitous rating scale. For example, in a typical attitude scale (sometimes called a Likert scale), respondents indicate whether they *agree very strongly, agree strongly, agree, neither agree nor disagree, disagree, disagree strongly,* or *disagree very strongly* with each attitude statement in the questionnaire. It is common to associate the numbers +3, +2, +1, 0, -1, -2, and -3 with those categories, respectively. The respondent's scale score is simply the sum of his scores for each item with a prior multiplication by -1 for items that run in the opposite "direction."

Unlike the representational approach to measurement, there are no strong checks of internal consistency that can be applied at the level of the individual entity and that could lead to a conclusion that the measurement model was violated. For example, knowing that someone *agreed strongly* with, say, attitude item a or that someone's total score on the attitude questionnaire is 36, provides no logical basis for predicting that person's response to, say, attitude item b. To be sure, given the way such scales are constructed, we would expect responses to items a and b to be correlated statistically across respondents. However, there is no way for a single individual to behave *inconsistently* with respect to the scale. For example, we can make the following claim without any knowledge of the content of items a and b: It is not inconsistent for someone to agree strongly with item a and to disagree very strongly with item b. We can perhaps say that such a situation is unlikely or statistically improbable, but there is no internal logic that renders such a pattern inconsistent for any pair of items.

If it is not possible to find a response pattern that is inconsistent with a particular non-representational scale, then it is impossible to falsify or invalidate the particular scale at the level of the individual entity. On the other hand, there are procedures for evaluating a scale or a measure at the aggregate level. For example, we might expect the responses to items a and b to be correlated positively on the whole, even though it is not logically inconsistent to agree with one and disagree with the other. The expectation of this aggregate correlation then provides a mechanism for the evaluation of the scale as a whole. Thus, lacking internal consistency checks, psychometric measurement relies on patterns of variances and covariances that can be evaluated only at the aggregate level and only probabilistically.

Axiomatic versus Psychometric Measurement

Some measurement theorists have suggested that non-representational measurement is an oxymoron. For example, Pfanzagl (1968) referred to measurement in the Likert scale tradition as "measurement by fiat." Such measurement theoriests believed that representational or axiomatic measurement as expounded in the three-volume Foundations of Measurement (Krantz, et al., 1971; Suppes, et al., 1989; and Luce, et al., 1990) would allow psychology to replace measurement by fiat with more defensible measurement procedures. However, more recently, Dawes (1994), who championed the representational approach to measurement in his chapter (Dawes and Smith, 1985) in the previous edition of the Handbook of Social Psychology, and Cliff (1992) have referred to axiomatic measurement theory as the revolution that "failed" outside psychophysics or that "never happened," respectively. Dawes and Cliff individually speculate about the reasons the revolution failed, including the difficulty of the underlying abstract mathematics, the lack of demonstrated empirical power, the apparently intractable problems of dealing with error, and the conflict with traditional research styles in psychology. Dawes (1994, p. 280) sums up, "the revolution outside psychophysics may have failed because the investigators too seldom managed to relate representational scale values . . . to anything else of importance."

While there are some success stories in psychophysics for representational measurement, successful applications of representational measurement in social psychology are difficult to find. Dawes (1994; see also Dawes & Smith, 1985) cites Coombs, Coombs, and McClelland (1975) as a successful application in which representational measurement procedures (a) resolved a theoretical problem that had stumped demographers trying to measure preferences for family size and composition and (b) created scales that had generality over a wide range of cultures. However, when these same techniques were applied to the study of other ferility-related topics such as the choice among contraceptive alternatives (Nickerson, McClelland, & Petersen, 1991), the results were less satisfactory.

While the revolution of representative measurement was sputtering, psychometric measurement was faring better in its ability to make predictions and making progress at getting its own house in order with respect to being able to test its measurement scales. We describe those developments, especially the use of confirmatory factor analysis, in later sections.

If the revolution represented by axiomatic measurement has failed, then why present it in this chapter? First, not everyone agrees that it has failed. Marley (1992, p. 96), for example, believes that the revolution is still in its early stages and that recent work on axomatic measurement represents "the foundations of measurement upon which future generations of theoreticians and experimentalists can build to test theories...." We suspect there is some wisdom in this prediction. Second, even if they are not used for actually constructing scales, axiomatic measurement models can often provide theoretical insights not available if we rely exclusively on traditional rating scale measures. We emphasize such insights in our quick tour of axiomatic methods. Third, we do think that there is a potential difficulty in the psychometric approach's exclusive reliance on statistical patterns of covariation to evaluate the adequacy of a measurement model. The problem arises because patterns of variance and covariance are used not only for this purpose but also to evaluate theoretical predictions about the relationships between different constructs. As a result, it can be difficult to know when covariation (or its absence) reflects the adequacy of a measurement model and when it reflects relationships between constructs predicted by theory. Suppose we find that two measures are highly correlated. When do we conclude that they are alternative measures of the same construct and when do we conclude that the relationship represents a theoretically meaningful association between different constructs?[1]

[1] This is similar to the well known conundrum in factor analysis concerning the "correct" number of factors to extract and rotate. Exploratory factor analysis is routinely used both to provide evidence for a measurement theory and to examine relationships between

This question would be less likely to arise if there were internal and logical consistency checks that could be conducted, independent of aggregate patterns of covariation, to assess the adequacy of the measures.

Alternatives to Numeric Scales

Before proceeding to the *how* of representational and psychometric measurement, we return to Steven's definition of measurement and his belief that it necessarily involves numeric representation. We think this is far too limiting. As we illustrate in the following examples, non-numeric scale values can often have important practical advantages in terms of communicating information about observations effectively.

Geometric Representations. Measurement and scaling are often associated with geometric representations of data. For example, scale values are sometimes thought of as points on a line as in Figure 1. Such geometric representations show the location of each entity on a single dimension or continuum and so such scales are often called "unidimensional."

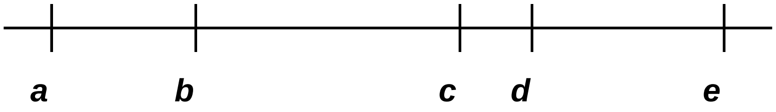


Figure 1.

Geometric, Unidimensional Representation of a Scale

The ubiquity of geometrical representations in measurement and scaling means that other geometrical concepts are often employed as convenient metaphors. One particularly pervasive metaphor is the geometrical concept of distance. Many social psychologists employ the concept of "psychological distance." The use of the adjective "psychological"

multiple constructs (i.e., oblique factors). Two individuals can look at the same correlation matrix and factor solution and concluded very different things about the number of constructs underlying the individual measures.

to modify "distance" indicates that it is not real distance that is meant but rather the psychologically effective distance between two entities. The distance between the location of two entities on a scale or dimension then represents this psychological or subjective distance. For example, if the scale in Figure 1 represents liberalness and the entities are attitude statements, then the psychological distance between items a and b is greater than the psychological distance between items b and c in terms of liberalness. Somewhat more general than psychological distance is the metaphor "functional distance," representing the functional effect of the entity, in terms of whatever dimension is being scaled.

Multidimensional Geometric Representations. Often we will measure more than one attribute of each entity so that a vector of scale values will be assigned. For example, in research using the semantic differential (Osgood, Suci, & Tannenbaum, 1957) we might want to assign three scale values to each noun to represent its evaluation (good-bad), its potency (strong-weak), and its activity (active-passive).

When more than one scale value is assigned to each entity, the entities can be represented as points in a multidimensional space with each dimension corresponding to a different type of scale value. For example, Figure 2 displays a two-dimensional scaling from Wish (1971) for 12 nations, based on their similarity ratings. Just as was the case for unidimensional measurement, we will sometimes be able to refer to the distance between points in multidimensional space as the functional distance between the corresponding entities.

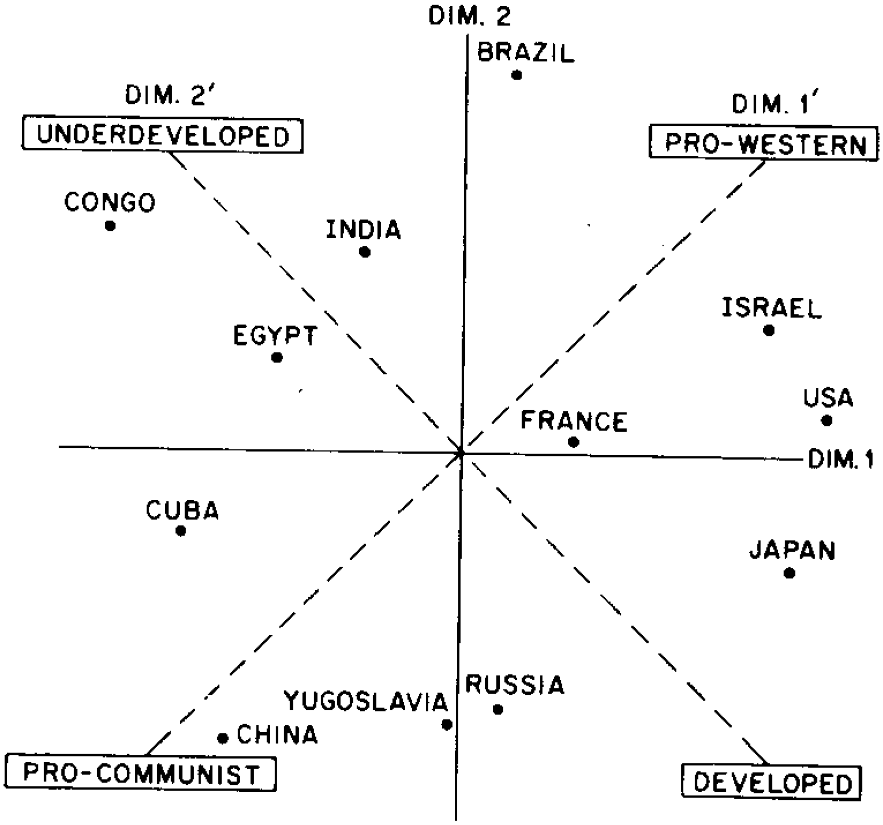


Figure 2.

Two-Dimensional Scaling of Similarity Judgments of 12 Nations

(from Kruskal & Wish, 1978, Figure 8, p. 32)

Non-Geometrical Representations. We also can use representations which are non-geometrical. For example, we might want to assign different shades of blue to be the scale values for the attitude statements with deeper shades of blue representing the more liberal statements and lighter shades the more conservative statements. Or we might want to assign a line drawing of a face to each census tract with the degree of the smile representing the average per capita income of the census tract (Chernoff, 1973).

Non-geometrical scale values would certainly not be very useful for some purposes—they would be difficult to input into statistical computer programs, for example. However, such scale values do have at least two important advantages. First, they can often be understood more readily. Many innovative methods for display information have been proposed (e.g., Cleveland, 1993b; Tufte, 1990), and there is active research and theoretical development about the intelligibility of visual displays and colored graphs (e.g., Cleveland, 1993a; Cleveland & McGill, 1987; Kosslyn, 1989; Simkin & Hastie, 1987; Shah & Carpenter, 1995; Tversky & Schiano, 1989; Wainer & Thissen, 1981) and on the presentation of complex multivariate tables by means of faces such as the one described above for census tracts (Chernoff & Rizvi, 1975). If the scales are to serve as input to a human computer instead of an electronic one, then nonnumerical scale values may well be preferred to confusing numerical tables.

A second advantage of nonnumerical scale values within the representational framework is that with a careful choice we can be sure that the scale values represent just the empirical relationship we observe. With numbers on the other hand, there can be some ambiguity unless we are very careful to specify just which properties of numbers are meant to apply. For example, we might assign 6 to attitude statement a and 2 to attitude statement b with the intention of only meaning that statement a is more liberal than statement b. Even if we are careful to state that only the ordering of the numbers is meaningful, some subsequent user of our scale might overlook that restriction and wonder whether a is three times more liberal than b because the scale value for a is three times that of b's. Such confusions are prevented if shades of color or some other nonnumerical scale values are assigned.

There is of course no reason why more than one nonnumerical scale value cannot be assigned to each entity. In fact, Chernoff's (1973) FACES were designed primarily for multivariate rather than univariate data. All that is required is a careful specification of what properties are to be represented by which characteristics. For example, Wainer and Theissen (1981) constructed the map of faces in Figure 3 to represent various observed properties of states that might be related to the quality of life by using the following assignment rules:

- Population: number of faces/state. (The number of faces is proportional to the log of the population.)
- Literacy Rate: size of the eyes (bigger = better)
- % HS Graduates: slant of the eyes (the more slanted the better).
- Life Expectancy: the length of the mouth (the longer the better).
- Homicide Rate: the width of the nose (the wider the nose the lower the homicide rate).
- Income: the curvature of the mouth (the bigger the smile the higher the income).
- Temperature: the shape of the face (the more like a peanut the warmer, the more like a football the colder).
- Longitude and Latitude: the position of the face on the coordinate axes of the paper.

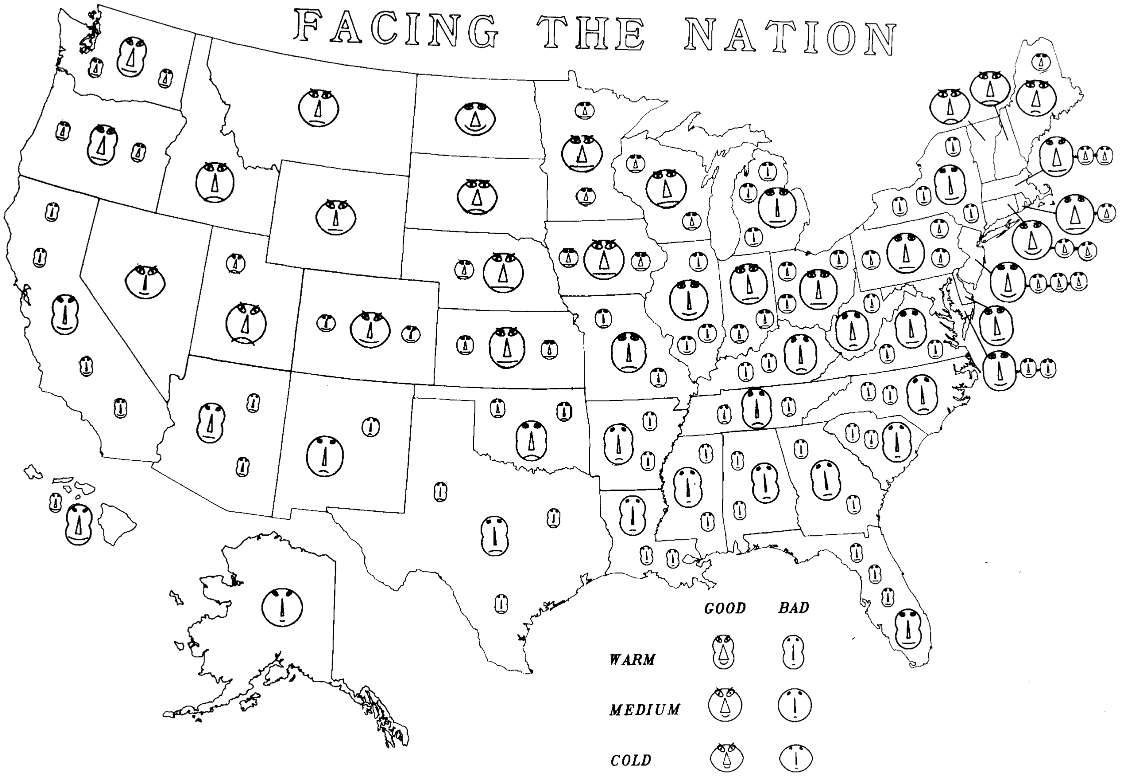


Figure 3.

Example of Chernoff's FACES for Displaying Multivariate Data

(from Wainer & Thissen, 1981, Figure 24, p. 230)

Figure 3 nicely illustrates the two advantages of non-numerical representations discussed above. First, a number of interesting comparisons and observations are almost immediate. For example, (a) it is striking that North and South Dakota are virtually identical on all the characteristics except for a dramatic difference in per capita income favoring North Dakota; (b) the greater population in the East is obvious; and (c) the faces in the South appear homogeneously grim reflecting a low quality of life with respect to these variables (see Wainer & Theissen, 1981, pp. 227--231, for a discussion of many other observations based upon visual inspection of this map). Second, note how Wainer and Theissen controlled the kinds of comparisons that could be made by their choice of features. For those characteristics for which a natural preference ordering is obvious (e.g., higher literacy is more desirable), a feature was chosen so that more of it was better (e.g., bigger eyes meant higher literacy) and so that the overall effect of more of the desirable characteristics was to create a happy, cheerful face. In contrast, for those characteristics for which there is not a natural preference ordering (e.g., temperature—some people want to ski and some want to swim), they chose a feature of the face (e.g., shape) which did not have a natural ordering and which did not make the face either more or less cheerful.

Axiomatic Measurement

Ordinal Measurement

We begin with ordinal measurement because it is the simplest example of axiomaic measurement. As such, it provides a good introduction to the basic ideas. Researchers in the social and behavioral sciences frequently want to assign scale values so that higher numbers represent more of the property being measured. The resulting scales are called ordinal scales because only the ordering of the numbers are important. A classical ordinal scale from outside the social sciences is the Moh Hardness scale used in geology. Rock a is harder than rock b if and only if a can scratch b, in which case it is assigned a higher numerical scale value. The "if and only if" means that if rock c has a higher scale value than rock d, then we can be assured that c will scratch d.

The hardness scale illustrates the key component for ordinal scaling: the existence of a well-defined empirical relationship (in this case, a scratches b) that is to be represented by the greater-than relationship of real numbers. For example, the entities might be a set of attitude statements, the property to be scaled might be degree of liberalness, and the empirical relationship might be a pairwise majority vote of a panel of 15 randomly selected students. We will use $\succeq$ to represent an empirical greater-than relationship between entities ("$a \succeq b$" can be read as "a dominates b" or "a is at least as great as b") and we will restrict the use of " ≥ " to comparisons between numerical scale values. For example, in the case of the attitude statements and the specified empirical comparison, $a \succeq b$ would mean that a majority of the panel voted that statement a was more liberal than statement b, and $s(a) \geq s(b)$ would mean that the numerical scale value assigned to a is higher than that assigned to b.

Axioms for Ordinal Measurement. In measurement theories, axioms state the empirical conditions which must be satisfied before it is possible to construct a scale. In effect, the axioms tell us the internal consistency checks which need to be tested. For the case of ordinal scaling, if the numerical relationship ≥ is to represent the empirical relationship $\succeq$, then the latter must behave exactly like the former. For example, for any two specific scale values it is always possible to determine either that they are tied or that one is larger than the other—either $s(a) \geq s(b)$ or $s(b) \geq s(a)$ or both (in which case $s(a) = s(b)$. So we should expect to observe this same property when we examine the empirical relationship $\succeq$. Thus, we have the following necessary axiom.

Connectedness. Either $a \succeq b$ or $b \succeq a$ or both.

Connectedness simply means that it must be possible to compare all the entities with one another. It is important to note that this does not rule out ties; both $a \succeq b$ and $b \succeq a$ may be true (in which case we will write $a \approx b$, meaning a and b are equivalent). Connectedness simply means that some decision can be made for every a and b.

Another basic ordering property of numbers and hence a property of our ordinal scale values is transitivity. If $s(a) \geq s(b)$ and $s(b) \geq s(c)$, then it must be true that $s(a) \geq s(c)$. For example, if 3 ≥ 2 and 2 ≥ 1, then of course 3 ≥ 1. Requiring this same property for the empirical relation $\succeq$ yields the corresponding axiom which follows.

Transitivity. If $a \succeq b$ and $b \succeq c$, then $a \succeq c$.

An empirical relationship $\succeq$ on a set of entities is said to be a *weak order* if and only if it satisfies the two properties of connectedness and transitivity. It is generally simple to test these two empirical properties in a given context. For example, if attitude statement a had been judged to be more liberal than attitude statement b and attitude statement b had been judged to be more liberal than attitude statement c, it would be easy to check whether or not statement a was judged to be more liberal than statement c as predicted by the transitivity axiom. Although these two axioms seem almost trivial, there is certainly no assurance that they will be satisfied for the sets of entities and empirical relationships listed in Table 1. For example, in the case of attitude statements, the panel might balk and refuse to compare two statements because they appeared to pertain to different topics; that would violate the connectedness axiom. Transitivity can also easily be violated. There is no assurance, as economists and political scientists have known since the publication of Arrow's (1951) famous Possibility Theorem, that majority pairwise votes by the attitude panel will produce a transitive ordering even if everyone votes transitively. For example, if five students think the order of liberalness is $a \succeq b \succeq c$, five think it is $c \succeq a \succeq b$, and five think it is $b \succeq c \succeq a$, then a "wins" over b 10 votes to 5 and b wins over c 10 votes to 5. By transitivity, it should be the case that a would win over c but instead it loses 5 votes to 10. If that were to occur, then it would mean that the transitivity axiom was violated and because transitivity is a necessary property for ordinal scale values that would mean in turn that no ordinal scale for those attitude statements could possibly be constructed using that empirical definition of "more liberal than." The importance of the connectedness and transitivity axioms is that it can be proved (see Krantz, Luce, Suppes, and Tversky, 1971, p. 15) that an ordinal scale is possible if and only if those two axioms are satisfied in the data. This is an important and surprisingly powerful result because it tells us that if we want to construct an ordinal scale, connectedness and transitivity are the only empirical properties that need be checked in the data. If those axioms are satisfied then we may proceed with the actual construction of an ordinal scale.

Constructing an Ordinal Scale. Testing the axioms and constructing an ordinal scale is very simple. The process is most readily understood in the context of a real example. Table 1 contains a subset of the data from a study by Clark (1982) which investigated preferences for residential location within the Milwaukee area. In this table, a 1 in row i and column j means that $i \succeq j$. Clark used the following empirical definition: If from among those movers who moved to either location i or j and who could have moved to either, a majority moved to location i, then location i is presumed to be more preferred than location j and a 1 is entered in the appropriate row and column. (See Clark, 1982, for more details.)

Table 1.

Subset of Residential Location Preference Data from Clark (1982)

		Locations										
		a	b	c	d	e	f	g	h	i	j	Count
	a	X	1	1	1	1	1	1	1	1	1	9
	b	0	X	1	0	0	1	0	1	1	0	4
	c	0	0	X	0	0	0	0	1	0	0	1
	d	0	1	1	X	1	1	0	1	1	1	7
Locations	e	0	1	1	1	X	1	1	1	1	1	8
	f	0	0	1	0	0	X	0	1	1	0	3
	g	0	1	1	1	0	1	X	1	1	1	7
	h	0	0	0	0	0	0	0	X	0	0	0
	i	0	0	1	0	0	1	0	1	X	0	3
	j	0	1	1	0	0	1	0	1	1	X	5

Note: A "1" in row *i*, column *j* means that location *i* was preferred to location *j* by at least 50 percent of the movers who could have moved to either location *i* and *j* and who did mover to either *i* or *j*.

The first step is to test the two axioms. Connectedness is easy—there is an entry in every cell of the data matrix so every comparison has been made. That is, either $a \succeq b$ and/or $b \succeq a$ for every pair of locations. Transitivity can be checked straigtforwardly by testing each subset of three locations in each possible ordering of the three. For example, $a \succeq e$ and $e \succeq h$ so we can check if $a \succeq h$ as it should be if transitivity is satisfied. In this case, it is. However, testing each individual transitivity separately would amount to 720 tests of transitivity in this instance. An easier and mathematically equivalent approach is to seek a permutation or reordering of the rows and columns so that the 1's form a triangular pattern above the diagonal. It is an easy exercise to verify that such a triangular pattern will result if and only if transitivity is satisfied. A useful heuristic is to sum the entries within each row. The top row in the triangle must have the greatest number of 1's so it should have the greatest count. Similarly, the bottom row should contain no 1's (i.e., it is not greater than anything else). The other rows should be ordered similarly between these two extremes according to their counts. The necessary counts are displayed in Table 2. These counts imply the order

$$a \succeq e \succeq g \succeq d \succeq j \succeq b \succeq i \succeq f \succeq c \succeq h$$

although we note that there are some ties which will make it difficult to obtain a perfect triangular pattern. The same data matrix with the rows and columns permutted according to the above order is displayed in Table 2.

Table 2.

The Data of Table 1 Rearranged

to an Approximate Triangular Pattern

		Locations										
		a	e	g	d	j	b	i	f	c	h	Count
	a	X	1	1	1	1	1	1	1	1	1	9
	e	0	X	1	1	1	1	1	1	1	1	8
	g	0	0	X	1	1	1	1	1	1	1	7
	d	0	1	0	X	1	1	1	1	1	1	7
Locations	j	0	0	0	0	X	1	1	1	1	1	5
	b	0	0	0	0	0	X	1	1	1	1	4
	i	0	0	0	0	0	0	X	1	1	1	3
	f	0	0	0	0	0	0	1	X	1	1	3
	c	0	0	0	0	0	0	0	0	X	1	1
	h	0	0	0	0	0	0	0	0	0	X	0

Any 1's below the diagonal in Table 2 indicate a potential violation of transitivity; there are two such instances involving the location pairs (d, e) and (f, i). Both of these result from equivalences; that is, both $d \succeq e$ and $e \succeq d$ so $d \approx e$, and both $f \succeq i$ and $i \succeq f$ so $f \approx i$. The (f, i) pair poses no problems because it just means that according to this measure locations f and i are functionally equivalent so that they can be combined into a single row and column. However, that is not the case for the (d, e) pair; locations d and e cannot be combined because the data indicate that location g is between them. This is a violation of transitivity because $g \succeq d$, $d \succeq e$, but g is not $\succeq e$. This means that an ordinal scale cannot be constructed for these data. If we want to construct a scale anyway, then one of the locations d, e, or g must be eliminated or one of the empirical observations changed. In general, whether it will be wise to do so depends on other information such as the reliability of the empirical comparison. A return to the raw data reveals that the observation that $d \succeq e$ was due to the fact that exactly 50 percent of the movers choose d and exactly 50 percent choose e. There would be no problem with transitivity had the proportion choosing d over e been 49 percent. While Clark does not provide sufficient information to conduct a statistical test, it is unlikely that the difference between 50 percent and 49 percent is statistically reliable. Hence, we will eliminate the observation for $d \succeq e$. The resulting matrix is displayed in Table 3. The scaling is now complete because the count for each row can be the scale value for that row's location.

Table 3.

The Data of Table 2 Rearranged

to an Approximate Triangular Pattern

		Locations									
		a	e	g	d	j	b	f,i	c	h	Count
	a	X	1	1	1	1	1	1	1	1	8
	e	0	X	1	1	1	1	1	1	1	7
	g	0	0	X	1	1	1	1	1	1	6
	d	0	1	0	X	1	1	1	1	1	5
Locations	j	0	0	0	0	X	1	1	1	1	4
	b	0	0	0	0	0	X	1	1	1	3
	f,i	0	0	0	0	0	0	X	1	1	2
	c	0	0	0	0	0	0	0	X	1	1
	h	0	0	0	0	0	0	0	0	X	0

Summary of Ordinal Scaling. We now have everything we need to construct ordinal scales for any entities which may be of interest to us. We have axioms or consistency checks which specify the empirical conditions that must be satisfied in order to construct a scale, and we have a scaling algorithm for actually constructing the scale. As will often be the case, the scaling algorithm is intimately linked with checking the properties of the representation theorem. Rearranging the data into a form that facilitates checking the axioms will often be equivalent to constructing the scale if it is possible to do so.

There are many instances in the social and behavioral sciences where an ordinal scale is all that is required and in some cases is to be preferred. A test of ordinal scaling is preferred, for example, when our goal is to reject a particular theory. If a substantive theory, which we would like to test, predicts at least an ordinal scaling of a set of entities, then finding that the data did not satisfy the axioms for ordinal scaling would constitute a powerful rejection of the theory. For example, Tversky (1969) considered alternative models of how people might make choices between entities described by several dimensions. Additive models, in which an overall evaluation is formed by adding together separate evaluations of the components of each entity, were rejected because those models predicted an ordinal scale, but the data showed clear violations of the transitivity axiom necessary for an ordinal representation. Measurement and scaling theories are just like any other theories so we always have more confidence in their rejection than in their confirmation.

In other situations where an actual scale is required, an ordinal scale may be sufficient. For example, if the task is to select the three best candidates from among a pool of applicants, then an ordinal scale of the applicants provides all the information that is necessary. Furthermore, in practice, ordinal scales are just as reasonable to be used in statistical analyses as most of the variables used by social psychologists. For example, it would be informative to correlate the ordinal scale from Table 3 with census tract information to identify possible bases for neighborhood preferences. The simplicity, transparency, and usefulness of ordinal scales suggest that they should be used much more in social psychology than they are.

Thurstone and Fechnerian Scaling

With an ordinal scale one, of course, does not know how large the differences are between adjacent entities. For example, although we know from our ordinal scale that location a is a more desirable location than e, we do know whether there is a large or small difference in preference between the two. Thurstone (1927) saw how to transfer ideas from psychophysics to the scaling of social stimuli. Psychophysicists had created scales by assuming that just noticeable differences were subjectively equal. Thurstone substituted equally-often noticed differences for just noticeable differences. That is, if in a number of trials, either within or across respondents, the empirical relation $a \succeq b$ was observed 85 percent of the time and the relation $c \succeq d$ was also observed 85 percent of the time, then Thurstone created scales by assuming the psychological distance between a and b equaled the psychological distance between c and d. This was a significant breakthrough because it meant that scaling could be accomplished without starting with physical measurements of the stimuli as required in psychophysical scaling. To construct scale values, Thurstone based a model on the normal distribution. Subsequent work in measurement theory has revealed that assumptions about a particular probability distribution are not crucial. We first present Thurstone's model because it is the most practical way to obtain scale values and then we consider the internal consistency checks implied by the notion that equally often noticed differences are equal.

Thurstone's Model. Thurstone argued that stimuli, whether they be lights in the psychophysicist's lab or attitude statements in a survey, never strike us in exactly the same way because of all the usual things that contribute to error in psychological processes. He assumed that on some undefined scale of psychological intensity in the head, each stimulus would have a normal distribution of possible ways in which it might strike the observer. We would expect the impression of a given stimulus to be near the mean on average, but the normal distribution implies that sometimes, but infrequently, that impression could be a considerable distance from the mean. We will let $\psi(a)$ be the particular impression experienced when stimulus a is presented; $\psi(a)$ is modeled as a normal random variable with mean μ_a and variance σ_a^2. When a judge is asked which, for example, of two attitude statements a or b is, say, more liberal, then, according to Thurstone, two impressions $\psi(a)$ and $\psi(b)$ are obtained. The judge then reports that "a is more liberal than b," which we record as $a \succeq b$. From these assumptions it is straight forward to derive what Thurstone referred to as the *law of comparative judgment*, although it really is just a theoretical model and not a law, namely,

$$s(a) - s(b) = z_{ab}\sqrt{\sigma_a^2 + \sigma_b^2 - r_{ab}\sigma_a\sigma_b}$$

where z_{ab} is the z-score corresponding to the probability of $a \succeq b$ and r_{ab} is the correlation between the respective normal distributions. Only z_{ab} is observed, so this model has far more parameters than observations. Simplifying assumptions are needed to reduce the number of parameters. Thurstone identified five cases corresponding to different sets of assumptions. By far the most commonly used is Case V, which makes the strong assumption that the underlying normal distributions are independent of one another so that $r_{ab} = 0$ and that they have the same variance, which can be fixed at 1 without loss. With these assumptions, the model simplifies to

$$s(a) - s(b) = z_{ab}$$

If all possible comparisons are made among *k* entities, then there are *k* scale values to be estimated from *k(k-1)* z-scores. Standard regression programs can provide the estimated scale values; it turns out, that the least-squares estimate for an entity is simply the average of all the z-scores in which that entity is involved.

We illustrate the construction of a Thurstone scale using data collected for this chapter. Forty-five students saw all possible pairs of six attitude statements and judged which statement in each pair they believed was the more liberal statement. The statements and the probability with which the row statement was judged more liberal than the column statement are listed in Table 4.

Table 4.

Affirmative Action Attitude Statements and Probability

that Row Item Was Judged More Liberal than Column Item

a In college admissions, at least 15% of all openings should be reserved for Blacks.

b Blacks should be given a special break in college admission decisions regardless of their qualifications.

c If two candidates are equally qualified for college admission, a slight preference should be given to the Black candidate.

d I'm all for affirmative action admissions, but that shouldn't mean that a qualified white candidate gets turned down in favor of a less qualified black candidate.

e Only merit, rather than any affirmative action considerations, should determine college admissions.

f Affirmative action programs in college admissions only result in underprepared students being admitted.

	a	b	c	d	e	f
a	—	.58	.73	.71	.78	.80
b	.42	—	.73	.76	.76	.84
c	.27	.27	—	.69	.71	.87
d	.29	.24	.31	—	.67	.73
e	.22	.24	.29	.33	—	.71
f	.20	.16	.13	.27	.29	—

Table 5 displays the probabilities of Table 4 converted to z-scores, with the average z-score in the last column representing the estimated scale value. In addition to information about the ordering of the items on liberalness, the Thurstone scale also provides information on the relative spacing of the attitude statements. This relative spacing is easier to see when points representing the items are arranged on a line as in Figure 4. Clearly, students did not judge much difference in the liberalness of items *a* and *b*, while they did judge a large difference between the liberalness of items *e* and *f*.

Table 5.

z-scores for Probabilities of Table 4.

	a	b	c	d	e	f	Average = Scale Value
a	0	.20	.62	.56	.76	.84	.50
b	-.20	0	.62	.69	.69	1.01	.47
c	-.62	-.62	0	.49	.56	1.11	.15
d	-.56	-.69	-.49	0	.43	.62	-.11
e	-.76	-.69	-.56	-.43	0	.56	-.31
f	-.84	-1.01	-1.11	-.62	-.56	0	-.69

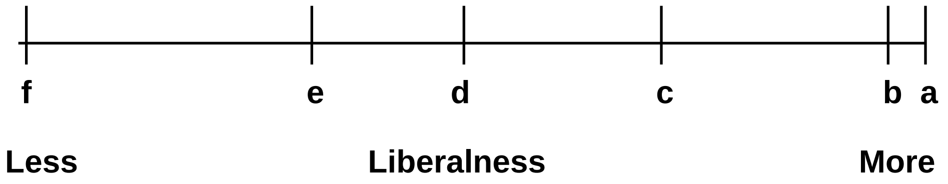


Figure 4.

Liberalness Scale Values Depicted on a Line

The Thurstone scaling model is indeed a model, so the scale values can be inverted to generate predictions. That is, once we have the scale values from Table 5, we can use the model equation

$$s(a) - s(b) = z_{ab}$$

to derive predicted z-scores, which in turn lead to predicted probabilities. For example, s(a) = .5 and s(c) = .15 so Z_{ac} =.5 - .15 = .35. Converting this z-score to a probability yields a prediction that the probability that a is judged more liberal than c ($a \succeq c$) is .63. All the predicted z-scores and probilities are displayed in Table 6.

Table 6.

Predicted z-scores and Probabilities Derived from Thurstone Scale Values of Table 5

Predicted z-scores

	a	b	c	d	e	f
a	0	.03	.35	.61	.81	1.2
b	-.03	0	.32	.58	.78	1.2
c	-.35	-.32	0	.27	.47	.84
d	-.61	-.58	-.27	0	.20	.57
e	-.81	-.78	-.47	-.20	0	.37
f	-1.2	-1.2	-.84	-.57	-.37	0

Predicted Probabilities

	a	b	c	d	e	f
a	—	.51	.63	.73	.79	.88
b	.49	—	.62	.72	.78	.88
c	.37	.38	—	.61	.68	.80
d	.27	.28	.39	—	.58	.72
e	.21	.22	.32	.42	—	.65
f	.12	.12	.20	.28	.35	—

The Thurstone scaling model can then be tested by comparing the observed and predicted probabilities. For example, the predicted probability of .63 for $a \succeq c$ was observed to be .73. Standard procedures can be used to compare the the observed and predicted probabilities. In this case, $\chi^2(10) = 15.84$, p = .10. Hence, the Thurstone scaling model is not rejected for these data. Had the model been rejected, this omnibus test would not tell us which aspect of the Thurstone model were incorrect: the specific Case V assumptions, the normal distribution assumption, and/or the basic idea that equally often noticed differences are subjectively equal.

Fechnerian Scaling. Since Thurstone's pioneering work, measurement theorists have made much progress in identifying the key properties of Thurstonian scaling. First, it was noted by Burke and Zinnes (1965) and Yellott (1977) that subsequent scaling models attributed to Bradley and Terry (1952), Luce (1959), and Dawkins (1969) make very similar predictions even though those models have very different assumptions, especially different implicit assumptions about the underlying probability distribution. This suggested that it should be possible to distill the more basic principles underlying Thurstone scaling and similar models. The essence of the idea that equally often noticed differences are equal can be presented by

$$P(a \succeq b) = F[s(a) - s(b)]$$

where *P* represents the probability of $a \succeq b$ and *F* is any montonoically-increasing function. For Thurstone scales, *F* is the cumulative normal probability function. Scales consistent with the above equation which allow *F* to be any increasing function are often referred to as generalized Thurstone scales or Fechnerian scales in honor of the source of the psychophysical scaling ideas that Thurstone extended. Although the Fechnerian scaling model represented by the above equation avoids making any assumptions about an underlying probability distribution, it makes surprisingly strong predictions. One such prediction is

Independence: $P(a \succeq c) \geq P(b \succeq c) \Leftrightarrow P(a \succeq d) \geq P(b \succeq d)$.

In other words, as Tversky and Russo (1969, p. 3) state, "if two stimuli [(a,b)] are ordered according to their choice probabilities relative to some fixed standard then, ... the ordering is independent of the particular standard." This implies that it should be possible to reorder the rows and columns of a matrix of probabilities so that the probabilitiies decrease within each column and increase within each row. Examining the rows and columns of the probabilities in Table 4 we see that this is generally, but not completely, true. Notably, if the order of the attitude statements on the scale is indeed a to f, then the greatest difference between scales values ought to be between s(a) and s(f) so that the largest probability in the matrix ought to be $P(a \succeq f)$, but in fact its probablility is only .80 which is less then the probability of .87 for $P(c \succeq f)$. Rearranging the row and columns so that c and f would be the furtherest apart leads to even more serious inconsistencies of the orderings within rows and columns. Hence, the ordering from a to f is the best ordering, but it violates a necessary condition for constructing a Fechnerian scale. Unfortunately, the Achilles heel of measurement theory models such as this is that it is difficult to decide whether the disprecancy between .80 and .87, which is not statistically significant, and the other discrepancies in Table 4 are large enough to invalidate the Fechnerian scaling model. Luce et al. (1990, p. xiii) admit, "The [planned] chapter on statistical methods was not written, largely becuase the development of statistical models for fundamental measurement turned out to be very difficult."

If independence and other related necessary conditions (e.g., see Michell, 1990, for details) are satisfied, then it is possible to solve the system of inequalities generated from

$$P(a \succeq b) \geq P(c \succeq d) \Leftrightarrow s(a) - s(b) \geq s(c) - s(d)$$

to provide rather precise estimates of the scale values. It is surprising how strongly the ordinal relationship constrain the possible scale values (McClelland & Coombs, 1975; Lehner & Noma, 1980, Roskam, 1992). However, the algorithms for solving the system of inequalities will only be successful if there is not a single violation of any of the necessary conditions.

Coombs' Unfolding Model

The Thurstone scaling model, when used with data collected from many individuals, assumes that everyone is viewing the entities from the same perspective. For example, the assumption for the affirmative action attitude statemens is that everyone, except for some random error, agrees on the ordering of the items with respect to liberalness. While that assumption may or may not be correct for liberalness judgments, it is certainly not correct for preferences. A very liberal person and a very conservative person might well agree on the ordering of the attitude statements from conservative to liberal, but their preference orderings might be exactly opposite. Coombs (1950, 1964) formalized such a situation in his unfolding model. In addition to scaling the items, the unfolding model also adds a scale value for each individual's "ideal" point on the same scale as the items. Figure 5 illustrates the unfolding model. The notion is that individuals, when confronted by a choice between two items, prefer the item which is closer to their ideal point. For example, in Figure 5, individual *i* 's ideal point is closest to item *b,* which would be his or her first choice, next closest is *a*, and so on; thus, *i* 's preference ordering would be

$$b \succeq a \succeq c \succeq d\,.$$

While for individual *j,* the preference ordering would be

$$c \succeq d \succeq b \succeq a\,.$$

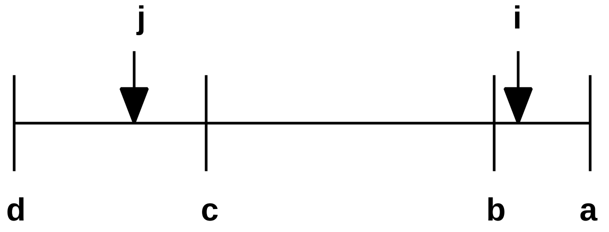


Figure 5.

Ideal Points for Individuals on the Same Scale as Items

Formally, we can state Coombs' unfolding model as

$$a \succeq_i b \;\Leftrightarrow\; |s(a) - s(i)| \leq |s(b) - s(i)|$$

which means that we would observe that individual *i* has a preference for *a* over *b*, if and only if the scale values for *a* and *i*'s ideal point are closer together than the scale values for *b* and *i*'s ideal point. Just as for Thurstonian and Fechnerian scaling, this model generates predictions that can be tested in the data. For example, if the ordering, but not necessarily the spacing, of the items is as in Figure 5, then the least preferred alternative for every individual must be either *a* or *d*. The model also implies less obvious predictions. For example, if there is an individual *i* for whom we observe

$$c \succeq b \succeq a \succeq d,$$

then we can infer that the distance between s(a) and s(b) is smaller than the distance between s(c) and s(d). If that is the case, then Coombs' model predicts that there should *not* be an individual with the preference ordering

$$b \succeq c \succeq d \succeq a$$

which would imply the opposite ordering of the distances. If a set of preference orderings is conistent with the Coombs' unfolding model, then it is possible to solve the resulting system of inequalities derived from the formal model (e.g., McClelland & Coombs, 1975). However, the procedures are sufficiently complex (cf., Mitchell, 1990) and the problems of error sufficiently difficult (cite probabilistic unfolding books) that it seems unlikely that many social psychologists will use the full unfolding model. Nevertheless, there are important insights to be gained by considering the model and even some simple and useful scaling techniques based on a partial unfolding model. We turn to those now.

Second Choices. When given a choice in a survey among several options, respondents seldom select an extreme alternative. Instead, many respondents, sometimes a majority, select the same alternative. In such cases, the choice provides little discrimination among the respondents. In those situations, Coombs unfolding model suggests it would be useful to ask respondents for their second and perhaps third choices. The second choice reveals the option that the individual is subjectivley next most closest to and will necessarily be towards one extreme or the other. For example, most young couples in the United States say they would prefer to have two children. Attempts to correlate preferences for family size with actual behavior fail because there is so little variance in the preference measure. However, additional variance in the predictor can easily be obtained by asking for second and third choices. For example, we would expect, on average, larger actual family sizes for couples whose preferences are $2 \succeq 3 \succeq 4$ than for couples who preferences are $2 \succeq 1 \succeq 0$, even though they share the same first choice. And in this context the ordering $2 \succeq 0 \succeq 1$ would reveal a strong bias against only children.

As another example of the usefulness of second choices, consider the Thurstone scaling of affirmative action statements from Figure 4. Figure 6 illustrates the response patterns predicted for the ordering and spacing of the items from the Thurstone scaling. For example, an individual whose ideal point on the affirmative action scale is to the right of the $\boldsymbol{dc}$ midpoint (who therefore is closer to $\boldsymbol{c}$ than $\boldsymbol{d}$), but to the left of the $\boldsymbol{db}$ midpoint (and therefore closer to $\boldsymbol{d}$ than $\boldsymbol{b}$) would have the first two choices $\boldsymbol{c}$, then $\boldsymbol{d}$. Note that six items have the potential of discriminating respondents into ten ordered categories from most favorable to least favorable towards affirmative action. Such discrimination is often sufficient for many social psychological studies.

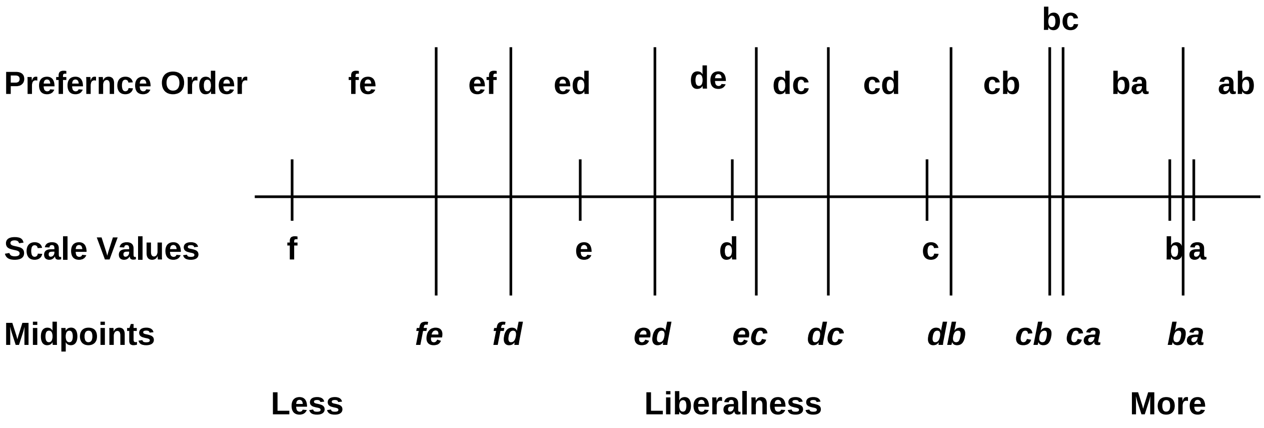


Figure 6.

Scale of Midpoints and Predicted Preferences for Order 2/6

Obtaining only the first and second choices does not provide enough information for the full set of internal consistency checks afforded by the complete unfolding model. However, there are some inconsistencies that can be detected. For example, if the affirmative action statements are ordered on liberalness as indicated by the Thurstone scaling, then there should be no respondents whose first two choices included non-adjacent items. Having a number of respondents with non-adjacent preferences would suggest that their preferences with respect to affirmative action programs could not be represented on a liberal/conservative dimension.

The same students who made the liberalness pair comparison judgments for the Thurstone scaling also indicated their first and second preferences among the six attitude statements. There were few instances of choice pairs including non-adjacent items. There were not enough preferences at the less liberal end of the scale to discriminate items ***e*** and ***f***, so they were combined. Figure 7 shows the frequency distribution of the preferences orderings. Note that most everyone include statement ***b*** ("Blacks should be given a special break in college admission decisions regardless of their qualifications") as either the first or second choice. The discrimination among respondents is then produced largely by the more liberal statement ***a*** or the less liberal statement ***c***. Using second choices to enhance differentiation among respondents and to provide light testing of preference models would be a useful addition to many studies in social psychology.

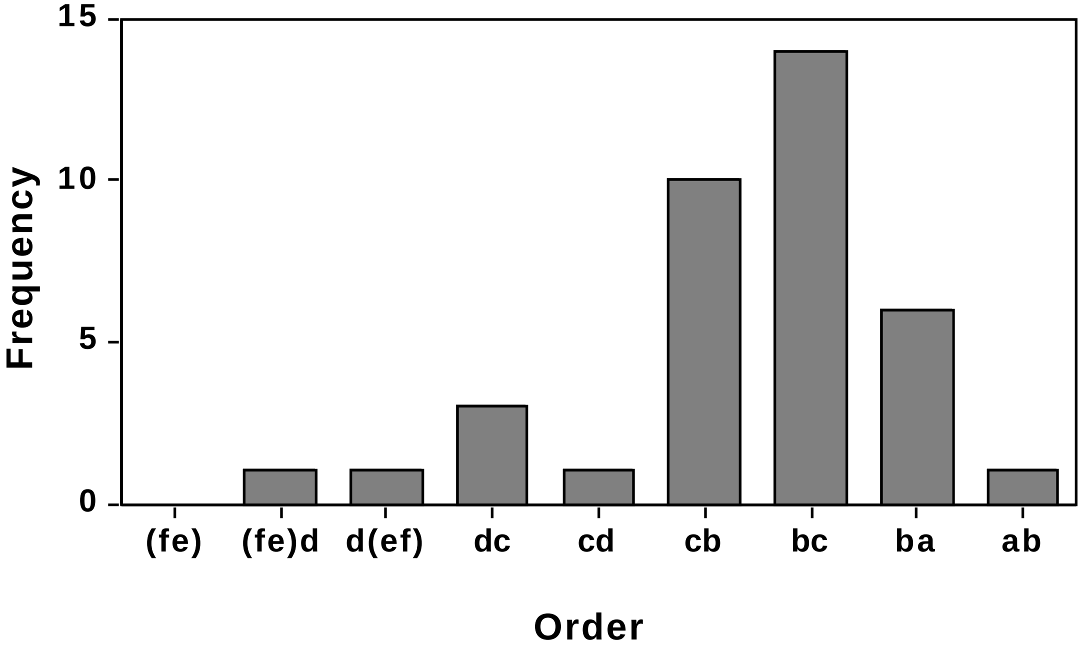


Figure 7.

Histogram of Preferences from Order 2/6 for Affirmative Action Items

Interpreting Survey Responses. The unfolding model, especially the notion that it is the midpoints between items that determine choices, is useful for understanding responses to simple questions often used in surveys. For example, suppose an opinion pollster asks a random sample of respondents how good a job the President is doing with response categories "very poor," "fairly poor", "neither good nor poor," "fairly good," and "very good." Assuming those response categories could be ordered along a continuum of approval, it would be the midpoints between items that would determine the responses. For example, according to the unfolding model, someone would select the category "fairly good" only if their opinion of how good a job the President was doing was closer to "fairly good" than "very good" and also closer to "fairly good" than "neither good nor poor." If a different category had been used for the top end, say "good" instead of "fairly good," then we would expect the proportion choosing "fairly good" to change because the midpoint between "fairly good" and the top category would have changed.

Failure to realize that it is the midpoints and not the labels of the response categories that determines choices can lead to confusion. For example, consider the hypothetical results in Table 7 from two polls A and B which asked how good a job the President is doing. The results seem quite different. Pollster A concludes that over half (50.5%) of the people are satisfied with the job the President is doing, while only 22% are definitely dissatisfied. Pollster B, on the other hand, concludes that only 35.5% are satisfied, while an almost equal proportion (34.5%) are dissatisfied. Note that in Poll A 37.5% chose the "fairly good" category, but only 17.5% selected that category in Poll B. Such disparate results might prompt us to question their sampling methods or even to suspect skullduggery. However, both polls are completely conistent with a common underlying distribution of ideal points. The two polls just divide the distribution differently. Figure 8 illustrates the consistency of the two polls. The line in Figure 8 represents the underlying dimension of approval of the job the President is doing. The midpoints for Poll A are depicted above the line and those for Poll B below the line. The proportion between each pair of midpoints (i.e., the proportion choosing each alternative as in Table 7) is displayed above and below the line, respectively, for Polls A and B. The combined midpoints divide the distribution of ideal points into the proportionate categories indicated on the line. The key difference in the two polls is that the middle or "neutral" cateogry in Poll A ("neither good nor bad") is viewed as more negative (in this hypothetical example) than the middle category in Poll B ("just OK"). With the position of of the middle category being more negative in Poll A, the midpoint between that middle label and the "fairly good" label is also necessarily more negative. Thus, some respondents (15%) who selected the middle category in Poll B would select "fairly good" instead of the middle category had they responded to Poll A. An unscrupulous pollster wanting the President to look good would select a "neutral" category label that was as negative as possible while one want the Presidnet to look bad would select a "netural" category that was as positive as possible.

Table 7

Hypothetical Results for Two Polls Asking the Same Question with Different Response Categories

Poll A		Poll B	
Category	Percent	Category	Percent
very poor	4.5	poor	12.0
fairly poor	17.5	fairly poor	22.5
neither good nor poor	27.5	just OK	30.0
fairly good	37.5	fairly good	17.5
very good	13.0	good	18.0

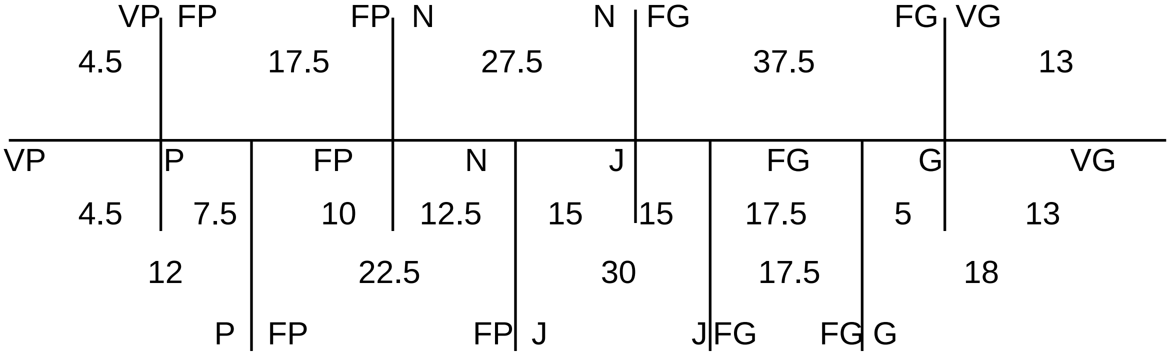


Figure 8.

Combined Scale and Midpoints for Two Polls

It is at first counterintuitive that using different middle categories can have such a dramatic impact on the proportion of respondents selecting more positive responses; however, it makes perfect sense in terms of the unfolding model. Finally, note that the dramatic difference in the proportions of each poll choosing "fairly good" is due to the large difference in neighborhing midpoints, both above and below. In Poll B, the "fairly good" category is squeezed by a closer middle category than in Poll A and a close more positive category ("good" rather than "very good" in Poll A).

Conflict and Social Choice. The unfolding model has interesting implications for the understanding of conflict, a topic of long-standing interest in social psychology. Here is one simple example. Suppose there are three individuals X, Y, and Z who have the following preferences for alternatives A, B, and C. (These alternatives might be political candidates or movies the group of three might go see or anything else they might be trying to agree about.)

$$\text{X: } a \succ b \succ c$$
$$\text{Y: } b \succ c \succ a$$
$$\text{Z: } c \succ a \succ b$$

If we ask this group whether they prefer a or b, there would be two votes for a and one for b, so we would conclude that for the group

$$a \succ_g b$$

And similarly, when we ask about b versus c, there would be two votes for b and only one for c, so we would conclude that for the group

$$b \succ_g c$$

Transitivity would then imply that

$$a \succ_g c$$

However, when we ask this group whether they prefer a or c, there would be two votres for c and only one for a. Thus, in fact,

$$c \succ_g a$$

which violates transitivity. A group with intransitive preferences (even though each individual has transitive preferneces) will have much difficulty making decisions and is likely to have a great deal of conflict among group members. The above illustration is the famous example that Condorcet (1785) used to demonstrate that majority voting is not guaranteed to be transitive (i.e., it is not guaranteed of producing even an ordinal scale).

Interestingly, a group will not face this sort of intransitive conflict if everyone's preferences can be represented by the unfolding model on a common unidimensional scale (Arrow, 1951). Further insight is gained by considering ordinal preference functions. Figure 9 depicts the preferences of the three individuals as ordinal functions over the three alternatives. The spacing of the alternatives in Figure 9 is only ordinal and the preferences are represented only as ranks. If the data meet the axioms for the unfolding model, then it is always possible to represent everyone's preferences as a single-peaked ordinal preference function. In Figure 9 it is clear for the alphabetical ordering of alternatives that individuals **X** and **Y** both have single-peaked preference functions, but **Z** has a single-dipped preference function. If we tried other orderings of the alternatives on the abcissa, we would find that at least one of these three individuals did not have a single-peaked preference function.

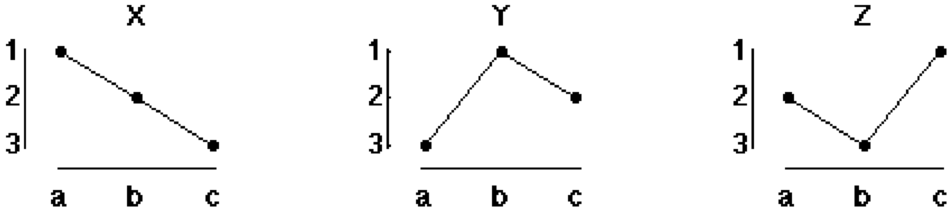


Figure 9.

Ordinal Preference Functions for Three Individudals

(Note: abcissa represents alternatives and ordinate is

rank-order preference with "1" being most preferred)

Presume for the moment that the correct ordering of the alternatives on an underlying dimension is indeed alphabetical so that it is **Z** who has a single-dipped instead of a single-peaked ordinal preference function. The problem is clearly that **Z** likes either extreme better than the middle. In other words, a prescription for creating group conflict is for some members of the group to prefer extreme alternative no matter in what direction those alternatives are extreme.

A famous example of group conflict created by preferring extreme alternatives occurred when the U.S. Congress considered the first major civil rights legislation in the 1960s. Let b represent the proposed civil rights legislation, let a represent an amended version of that legislation with many strengthening provisions added, and, finally, let c represent the status quo (no new legislation). The preferences of liberals corresponded to those of **X** above (i.e., they wanted the strongest possible legislation), and the preferences of moderates corresponded to those of **Y** (i.e., they wanted some legislation, but not if it were too strong). The Southern Democrats did not want any new legislation so their preferences were

$$c \succ b \succ a$$

However, the Southern Democrats realized that if they voted in this way, then the proposed legislation would pass. So, in order to produce conflict in the group decision, they *pretended* to have preferences corresponding to **Z** above. That is, they pretended to prefer either the status quo or a strongly amended bill to the the more moderate proposed legislation. In terms of the unfolding model, they were pretending to have preferences that could not be represented on the same unidimensional scale as those of the other two groups. In particular, they joined with the liberals to vote for strong amendments to the proposed legislation in the hope that the bill would become so radical that it would no longer appeal to the moderates. This plan was foiled when the liberals realized that voting according to their true preferences would produce a bill that was too strong to pass. So, they *pretended* that their preferences were

$$b \succ a \succ c$$

The possibility of group intransitivity was then eliminated because the ture preferences of the moderates, the pretend preferences of the Souther Democrats, and the pretend preferences of the liberals could all be represented in terms of the same unfolding model or, equivalently, as single-peaked preference functions over the alternatives when they are ordered on the abcissa in the same order as either the pretended liberal preference ordering. In other words, the pretend preferences again put liberals and Souther Democrats at opposite poles of a common unidimensional scale.

In the end, the civil rights legislation passed because so many congresspersons misrepresented their preferences! However, it is simple to understand what happened in the context of an unfolding scaling model for preferences. Coombs and Avrunin (19??) develop a number of other implications about conflict from the unfolding model and related ideas. Thus, even if representative or axiomatic models are not used to construct scales, they still are often useful for understanding social phenomenon.

Multidimensional Unfolding. So far we have assumed that the entities can be arrayed on a single dimension. There is of course no reason that entities cannot be distinguished in terms of multiple dimensions. A multidimensional unfolding model then applies. Figure 10 illustrates the location of 4 entities in a two-dimensional configuration. The lines divide the plane midway between each pair of entities. Again, the presumption is that people will prefer those entities that are closest to their ideal points. Thus, the lines separate the plane into profane regions; anyone with an ideal point within the same region should have the same preference ordering. The resulting preference orderings for each region are indicated in Figure 10.

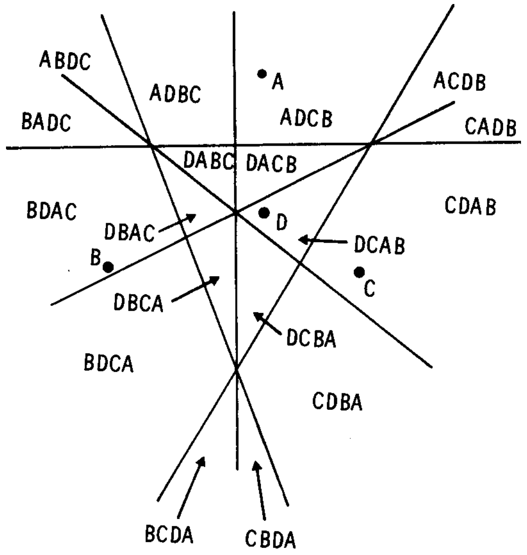


Figure 10.

Illustration of Multidimensional Unfolding

of Four Items in Two Dimensions

(Figure 12.5 from van der Ven, 1980)

Recovering the configuration—determining the number of dimensions and locating each entity on those dimensions—solely from knowing which preference patterns occur is forebiddingly difficult (see Bennet & Hays, 1960; Hays & Bennet, 1961; Coombs, 1964) and little is known about the necessary and sufficient conditions for multidimensional unfolding. In practice, multidimensional scaling programs, which conduct iterative searches to maximize some goodness-of-fit measure (Young, 1984), are used to find a multidimensional configuration. At the end of the chapter, we explore some implications of the multidimensional unfolding model in terms of the psychometric approach.

Conjoint Measurement

The most important representative measurement is additive conjoint measurement. Indeed, Fechnerian scaling, Coombs unfolding, and many other representative measures are in some sense special cases of conjoint measurement (Michell, 1990). We do not have space within a chapter to present anything close to a complete treatment of this important topic. Instead, we present a brief explanation of the major ideas and again focus on broader implications of those ideas, implications that apply whether or not one is using conjoint measurement. Accessible and more complete coverage is available elsewhere (e.g., Michell, 1990; van der Ven, 1980).

The key idea underlying conjoint measurement is to measure one variable against another, and vice-versa. For example, in a psychophysical context, suppose that we were able to establish that a person felt equally uncomfortable when the temperature was 80° with 93% humidity as when the temperature was 90° with 75% humidity. Then we could infer that increasing the temperature from 80° to 90° is exactly compensated psychologically by decreasing the humidity from 93% to 75%. In other words, the change in scale values for the temperatures must exactly equal the change in scale values for the humidity levels. Then, if we observe that a person felt equally uncomfortable at the combination (80°, 75%) as at (90°, 51%), we would know that the change from 80° to 90° is also exactly compensated by decreasing the humidity from 75% to 51%. This establishes that the difference in scale value between 75% and 93% humidity is equal to the difference in scale value between 51% and 75%. That is,

$$H(93\%) - H(75\%) = H(75\%) - H(51\%)$$

where H represents the scale for humidity. By equating intervals it seems obvious, and the measurement theorems of conjoint measurement confirm this (Luce & Tukey, 1964; Krantz et al., 1971), that we are building an interval[2] scale from such equivalences. If we let H(51%) = 1 and H(75%) = 2, in order to fix the origin and unit for the interval scale, then H(93%) must equal 3. Also, we know that

$$T(90°) - T(80°) = H(93\%) - H(75\%) = 1$$

[2] An interval scale is one in which the ratio of intervals remains constant under any permissible change in the scale values. In this case,

$$\frac{H(93\%) - H(75\%)}{H(75\%) - H(51\%)} = 1$$

Clearly, adding (or subtracting) a constant from each value of H() would not change this ratio and neither would multiplying each value of H() by a positive constant. Hence, such manipulations are permissible for interval scales. The additive constant determines (or is determined by the choice of) the origin and the multiplicative constant determines (or is determined by the choice of) the unit of measurement.

This is the conjoint part; once the unit of measurement is chosen for one scale, the same unit of measurement applies to the other scale. Scale construction would continue by finding the temperature that would make the equivalence (80°,75%) ~ (?,93%), in terms of uncomfortableness. Let's say that temperature were 72°; then the equivalence would establish that

$$T(80°) - T(72°) = T(90°) - T(80°) = 1$$

which would put us on the road to developing an interval scale for subjective temperature, an interval scale that is linked to the interval scale for subjective humidity. More importantly, we would also have the testable prediction that (80°,51%) ~ (72°,75%).

Seldom do social psychologists use variables like temperature and humidity that are quantitative and infinitely divisible in practice, nor are equivalences practical. However, this is not a difficulty because it is possible to utilize conjoint measurement when only ordinal information is available and for arbitrary entities. Let $A = \{a, b, c, \ldots\}$ be one set of entities and let $P = \{p, q, r, \ldots\}$ be another set. For example, A and P might be sets of attitude statements, one set about affirmative action and another set about abortion. We might ask respondents to order pairs of statements according to how liberal they would judge a person to be who endorsed both statements. Or, we could ask respondents to order the pairs of statements according to how favorable they would be to a political candidate who made both statements.

Conjoint measurement specifies the conditions that must be satisfied for an ordering of all pairs (a, p) to be represented by the addition of the respective scale values. That is,

$$(a, p) \succeq (b, q) \quad \Leftrightarrow \quad f(a) + g(p) \geq f(b) + g(q)$$

where f() and g() are the respective scale values for the two sets of entities. Necessary conditions that can be tested in the data are easily derivable.

Independence. $(a, p) \succeq (a, q) \Leftrightarrow (b, p) \succeq (b, q)$

and

$$(a, p) \succeq (b, p) \Leftrightarrow (a, q) \succeq (b, q)$$

Simply, if fixing the first component at some entity a implies that $p \succeq q$, then fixing the first component at any other entity b must also yield $p \succeq q$. Similarly, the ordering of entities from the first set must be the same no matter at which entity the second component is fixed. Figure 11 depicts the independence condition graphically. The condition requires the ordering of the columns within rows to be the same, regardless of row and the ordering of the rows within columns to be the same, regardless of column.

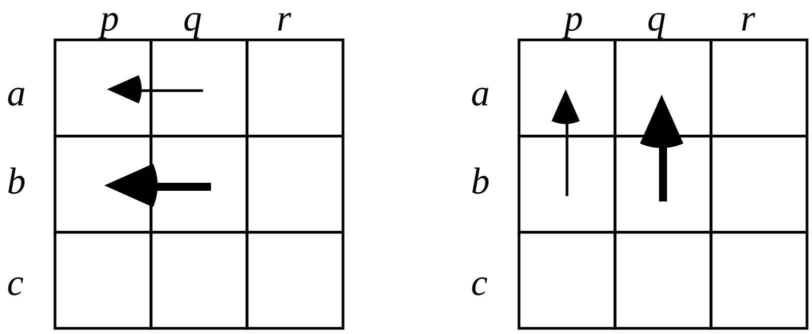


Figure 11.

Illustration of Independence Condition.

(The arrow tip represents the ordinally greater cell;

the small arrow implies the large arrow)

Double Cancellation.

If $(b, p) \succeq (a, q)$ and $(c, q) \succeq (b, r)$, then $(c, p) \succeq (a, r)$

The unusual name and the at first non-intuitive condition is best understood from its simple derivation. First, restate the two conditions in terms of the additive representation of the scale values:

$$(b, p) \succeq (a, q) \quad \Leftrightarrow \quad f(b) + g(p) \geq f(a) + g(q)$$

$$(c, q) \succeq (b, r) \quad \Leftrightarrow \quad f(c) + g(q) \geq f(b) + g(r)$$

Then, add the two inequalities to obtain:

$$\boldsymbol{f(b)} + g(p) + f(c) + \boldsymbol{g(q)} \geq f(a) + \boldsymbol{g(q)} + \boldsymbol{f(b)} + g(r)$$

Canceling the common terms (bolded) on both sides of the inequality (the "double cancellation"), yields the conclusion:

$$g(p)+f(c)\geq f(a)+g(r) \quad \Leftrightarrow \quad (p,c)\succeq(a,r)$$

Figure 12 depicts the double cancellation condition graphically.

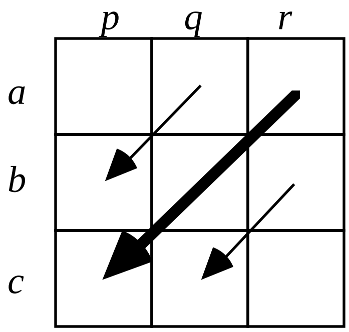


Figure 12

Illustration of Double Cancellation

(The two small arrows imply the large arrow)

With more than three entities in a set, there are tests of higher-order cancellation (triple, quadruple, etc.). However, the antecedents for those conditions become so complicated that the higher-order cancellation conditions are seldom violated. In practice, if independence and double cancellation are satisfied, one then tries to solve the system of inequalities generated by the ordering of the pairs (McClelland & Coombs, 1975; Lehner & Noma, 1980, Roskam, 1992). A solution results, producing scale values, if and only if all the necessary conditions are satisfied. If system of inequalities does not have a solution, then one can look for the violations of the higher-order cancellation conditions. Alternatively, some researchers use iterative algorithms such as MONANOVA (Kruskal, 1965), ADDALS (de Leeuw, Young, & Takane, 1976), PROC TRANSREG (SAS Institute, 1990), and generalized additive models (Hastie & Tibshirani 1990; Hastie 1992) in S+ to find a best-fitting additive solution. Such methods have been popular in marketing

(Green & Srinivasan, 1990; Wittink & Cattin, 1989). Anderson (1981, 1982, 1991) and colleagues adapt analysis-of-variance techniques for essentially the same purpose.

As with other representational measures, there are few successful applications of conjoint measurement in social psychology. In most cases, the axioms are violated so that scales cannot be successfully constructed. A particularly pernicious type of violation is illustrated by Tversky, Sattath, and Slovic (1988) who show that choice and matching (setting equivalences) can produce different orders. An axiomatic analysis was key to reaching that conclusion.

Coombs, Coombs, and McClelland (1975) provide a instructive example of the successful application of conjoint measurement in social psychology. They modeled the family composition preferences (i.e., the desired number of boys and girls). The additive model in terms of the number of boys and girls,

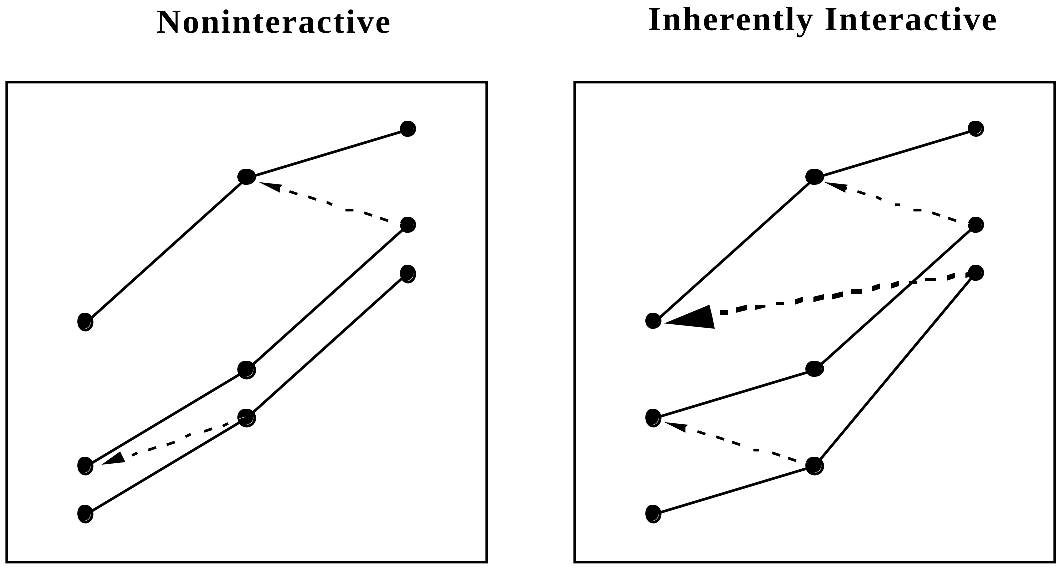
$$\mathrm{Pref}[B,G] = f(B) + g(G),$$

where B and G represent the number of boys and girls, respectively, was unambiguously rejected by substantial violations of independence in data from several different countries. Clearly, and not surprisingly, preferences for the ideal number of boys depended on the number of girls in the family, and vice-versa. However, most all violations of independence and double cancellation disappeared when the model was reëxpressed as

$$\mathrm{Pref}[T,D] = f(T) + g(D)$$

where T = B + G and D = B - G. Although it seemed reasonable to try to model preferences in terms of B and G, the obvious way in which family compositions vary, the conjoint measurement analysis revealed that total number of children and the difference in the number of boys and girls were the appropriate psychological variables.

Interpreting Interactions. Conjoint measurement has implications not just for measurement in social psychology, but also for data analysis in general. Conjoint measurement distinguishes between noninteractive models and inherently interactive models (Michell, 1990). Figure 13 illustrates both types of models in terms of graphs of data that might be obtained in a two-way design. In neither graph are the lines parallel; the question is whether there might be a monotonic transformation (i.e., one that preserves order) of the response variable that would make the lines parallel, and hence noninteractive[3]. If a monotonic transformation would make the lines parallel, then any conclusions about an interaction would depend on the assumption that the response variable had been measured on an interval scale. However, if no monotonic transformation could be found to make the lines parallel, then the interaction is inherent. Conjoint measurement provides a definitive answer as to whether such a transformation exists.



Noninteractive

Inherently Interactive

Figure 13

Illustrations of Noninteractive and Inherently Interactive Data

(Note: Small arrows indicate antecedents of double cancellation test;

if both small arrows point up or both point down,

then the large arrow must point in the same direction.

[3] Note that if all values are positive, then the multiplicative model

$$f(a)g(p)$$

is noninteractive because taking logs (a monotonic transformation) yields the additive model:

$$\log[f(a)g(p)] = \log[f(a)] + \log[g(p)]$$

The data values for the noninteractive model, the left panel of Figure 13, satisfy the axioms of additive conjoint measurement. Independence is clearly satisfied (each line is monotonically increasing and no lines intersect) and double cancellation is satisfied vacuously because the antecedent condition is not true. Thus, there exists a monotonic transformation of the response variable that would make the lines parallel. In contrast, the data values for the inherently interactive model, the right panel of Figure 13, fail the axioms of additive conjoint measurement. Although independence is satisfied, double cancellation is violated (if the two small arrows both point up or both point down, then the large arrow must point in the same direction). Thus, there is no monotonic transformation that could make the lines parallel; the interaction is essential or inherent and a claim for the interaction does not depend on having measured the response on an interval scale.

An implicit assumption underlying additive conjoint measurement is that noninteractive models are preferred. Luce (1995, p. 21) states this explicitly:

> ...evidence of interactions is usually a signal of trouble.... All too often, in my opinion, the interactions are treated as a finding and not as evidence of a lack of understanding of the combining rule for measures of the independent variables.

A lot of current work in social psychology emphasizes interactions.
Perhaps many of these interactions are not inherent and would disappear with appropriate scale transformations. Hence, the perspective of conjoint measurement suggests it might be more fruitful to identify fundamental variables that combine noninteractively to influence social behavior.

Psychometric Measurement

Representational or axiomatic measurement has much to recommend it, since the quality of measurement or the scalability of the measures can be examined through the sort of internal consistency checks we have reviewed. Nevertheless, most measurement in social psychology has not adopted the representational approach in spite of repeated admonitions to do so (e.g., Dawes & Smith, 1985). Most measurement in social psychology consists of questionnaire and observational measures whose validity is established not by a set of axiomatic consistency tests but rather by the observed patterns of variances and covariances that they display.

In general, we want to measure the attributes of some set of objects. Most typically in social psychological research, these objects consist of subjects and groups of subjects. And the set of attributes includes their traits, dispositions, attitudes, preferences, aptitudes, performances, and so forth. We collect data with the goal of ordering subjects on the measured variable or variables so that this observed ordering is the same as the unknown ordering on the true or latent construct that we wish to measure. A variable is said to possess high construct validity if these two orderings are highly similar.

The data that we use to order subjects on variables comes from a variety of sources and each of these sources makes certain assumptions about the subjects. Most typically, we ask subjects to report on themselves. They indicate their preferences, their values, their attributes by responding to questionnaire and interview measures. So the attitude researcher directly asks subjects for their evaluation of the attitude object. The researcher who studies stereotypes and intergroup relations asks subjects to provide descriptions of how they perceive various target groups. Someone interested in intimate relationships asks couples to report on the quality of their relationships. All of these ways of gathering data make the strong assumptions that a) subjects have access to the psychological property that the researcher wishes to measure, and b) subjects are willing to report that property.

Other, less direct measures assume that subjects have access to that which we wish to measure but may be unwilling to provide accurate self-reports. Thus, these approaches attempt to minimize or otherwise overcome self-presentational concerns, typically through deception. Examples include: a) The bogus-pipeline approach to attitude measurement (Aguinis, Pierce, & Quigley, 1995; Jones & Sigall, 1971) in which subjects are led to believe that they might as well respond truthfully since the researcher has direct access to his or her emotional response to the stimulus object; and b) Randomized responding procedures, again in attitude measurement (Dawes & Morre, 1979; Greenberg, Abdula, Simmons, & Horvitz, 1969; Warner, 1965), in which the subject knows that the researcher is unaware of the content of specific questions being asked but, unbeknown to the subject, the researcher is nevertheless able to infer attitudes from patterns of responses across multiple questions that have been sampled with some known probability from a universe of questions.

Other measurement approaches assume that there exists a psychological attribute to measure but that subjects may not have access to it and are unable, for whatever reason, to provide accurate self-reports. Unobtrusive observations of behavior, and inferences about attitudes and attributes from those behaviors, fall into this class of data collection approaches (Webb, Campbell, Schwartz, & Sechrest, 1966). Thus to measure ethnocentrism, we observe how closely subjects are willing to sit to an outgroup member (Macrae, Bodenhausen, Milne, & Jetten, 1994), to measure interpersonal intimacy, we might examine whether individuals addopt similar and synchronous nonverbal gestures (Bernieri & Rosenthal, 1991), and to measure alcohol consumption, we count bottles and cans in garbage containers (Webb et al, 1966). Measures in this category may also assess responses over which subjects presumably have little or no control, assuming that these responses are indicators of underlying psychological attributes or states. Thus, interest or attention might be measured by pupil dilation (Petty & Cacioppo, 1983; Woodmansee, 1970), arousal by galvanic skin response (Cook & Selltiz, 1964; Rankin & Campbell, 1955), and attraction or revulsion by minute movements in relevant facial muscles (Cacioppo & Petty; 1979; Petty & Cacioppo, 1983). Response latency measures, widely used in social cognition research (e.g., Dovidio, Evans, & Tyler, 1986; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Wittenbrink, Judd, & Park, in press), also fall into this class.

Regardless of the assumptions that one makes about the willingness and ability of subjects to provide accurate reports, the validity of all of these measures must ultimately be demonstrated by confirmation of expected covariance patterns in the data. Appeals to the facial validity of a variable ("It certainly seems to ask appropriate questions given our measurement goals.") are insufficient. Interestingly, in our opinion, such appeals seem somewhat more likely with less direct measurements. Thus, for instance, one may be more likely to encounter validity and reliability information (in the form of appropriate correlations and covariances) in the case of questionnaire measures than in the case of response latency data. In both cases, of course, the validity and reliability information is necessary.

The central question in measurement concerns the construct validity of measured variables. Each variable is assumed to be only an imperfect indicator of the underlying theoretical construct that one wishes to measure (Cook & Campbell, 1979; Cronbach & Meehl, 1955). According to traditional test theory, measures are imperfect because they necessarily contain some degree of random error or unreliability. According to more recent and comprehensive approaches, measures are imperfect indicators of constructs of interest not only because of random errors of measurement but also because they measure unintended constructs, systematic error, or what we will call constructs of disinterest. Thus, the latent reality of any measured variable is that the scores reflect three potentially distinguishable things: the construct that we would like to be measuring, a variety of constructs of disinterest that we would rather not be measuring, and random error or unreliability.

One argues for the construct validity of a variable by showing that its observed covariances and correlations with other variables provide evidence of 1) convergent validity, 2) discriminant validity, and 3) reliability. These three components of construct validity can be mapped on to the three categories of the latent reality underlying the variable. Thus convergent validity amounts to the demonstration that the variable reflects the construct of interest. Discriminant validity is the demonstration that constructs of disinterest are not in fact being measured. And reliability amounts to the demonstration that random errors of measurement are not large. Again, each of these demonstrations relies solely upon observed variance/covariance patterns rather than the sort of internal consistency checks found in the axiomatic approach to measurement.

Only relatively recently in the psychometric tradition has it been recognized that the central measurement question is that of construct validity (Messick, 1989; 1995). If one surveys older measurement texts (e.g., Anastasi, 1961; Cronbach, 1964) or testing and measurement standards as set forth by the American Psychological Association (1966), one finds references to additional criteria that must be satisfied to establish measurement quality. Thus, in addition to construct validity, these sources identify content validity, predictive validity, and concurrent validity as additional considerations that must be satisfied. More recently, however, the literature has recognized that these different forms of validity really can be subsumed under the general heading of construct validity and that the differences among them amount to differences in the sort of evidence used to establish construct validity. Thus, for instance, predictive validity is established by showing that a given measure predicts subsequent standing on other criterion measures. Concurrent validity examines covariances with other criterion measures taken simultaneously. We concur with others in the field (Cronbach, 1984; Messick, 1989; 1995) that these simply amount to alternative sources of evidence from which one gathers information about the construct validity of a variable. Covariance patterns with a variety of different other variables can be informative for different aspects of construct validity. The fundamental question is whether the variable measures what we want it to measure and not what we don't want it to measure.

To know where to look for evidence of construct validity requires that the construct that one wants to measure be theoretically defined. Thus, as we argued earlier, measurement once again presupposes theory. It is from theory that expectations are derived about observed patterns of variance and covariance that the measured variable ought to exhibit. The construct that is measured is embedded in a theoretically defined "nomological net" (Cronbach & Meehl, 1955) and the pattern of covariances that one expects with other measured variables is the manifestation of this net.

Finally, this approach to measurement relies upon patterns of linear association among variables. Linear associations are not, however, scale-invariant. Non-linear transformations of variables (e.g., log or inverse tranformations) can affect the pattern of covariances that are observed. Given that the psychological metric of a variable may not be the same as the measured metric, one needs to attend to whether the appropriate metric has been used in examining the construct validity of a variable. Most importantly, one needs to be sensitive to reasonable transformations and their effects on observed variance / covariance patterns. Fortunately, nonlinear but monotonic transformations tend to have small effects on observed correlations. But there do arise occasions when one ought to be invoking a strongly nonlinear and nonmonotonic model of association, such that more moderate scores on one variable ought to be associated with higher scores on another, for instance. Unless one realizes that one is confronted this situation, traditional approaches to construct validity are unlikely to yield informative results.

This section of the chapter is organized as follows. First, we go through the classic psychometric approach to reliability estimation. We then turn to a more general approach to the same issue involving the computation of components of variance and intraclass correlations. We then turn to a consideration of convergent and discriminant validity. First we consider the general approach to these topics and demonstate the utility of the multitrait - multimethod correlation matrix (Campbell & Fiske, 1959). Then we return to a discussion of variance components in multi-faceted designs, where systematic error components are manipulated and their contributions estimated (Cronbach, Glesser, Nanda, & Rajaratnam, 1972). Finally, we discuss in some detail a more general approach to construct validity that relies on confirmatory factor analysis. We illustrate its utility in examining multitrait - multimethod matrices and show its relationship to the variance components approach.

Classic Test Theory and Reliability Estimation

As we have already noted, the classic approach to measurement assumes that each observed variable has two underlying components: true score and random error:

$$X_i = T_i + E_i$$

At a later point, we will expand this model to include systematic error or constructs of disinterest, but for present purposes we can use this simpler model to motivate traditional approaches to reliability estimation. Recall that reliability is a component of construct validity. In theory it tells us about the relative amount of random errors of measurement in a variable. As such, it places an upper limit on the construct validity of the variable, since the more random error in it, the less it can be measuring the construct of interest.

From the above equation and the assumption that errors of measurement are random and therefore uncorrelated with true scores, it follows that the variance of the measured variable can be broken into two components: variance due to the true score and error variance:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

From this, we can define the reliability of a measured variable as the proportion of its total variance that is true score variance:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

The square root of the reliability can be shown to equal the correlation between the measured variable and the true score.

To estimate the reliability of a variable, we begin by assuming that we have what are called "parallel forms" of the variable. That is, we have two measures, X_1 and X_2, each of which contains the identical true score variation and each of which contains random errors of measurement to the same degree. Note that since the errors are random, we are not assuming identical errors in the two variables. Rather we are assuming equal error variances. Under these assumptions, it can be shown that the correlation between the two variables estimates the reliability of each:

$$r_{X_1X_2} = \rho_{X_1X_1} = \rho_{X_2X_2}$$

With k parallel measures, each of the k(k-1)/2 bivariate correlations estimates this same reliability. Additionally, since the errors in each measure are random, they tend to cancel each other out when we form a single measure by summing scores on the set of k measures. The extent to which this is true is given by the Spearman - Brown Prophecy Formula for the reliability of the sum of k parallel measures:

$$\rho_{SUM} = \frac{kr_{ij}}{1 + (k-1)r_{ij}}$$

where r_{ij} is the average of the k(k-1)/2 bivariate correlations between pairs of measures.

The assumptions of parallel measures are strong. One might want, for instance, to allow unequal error variances in the various measures, but still estimate the reliability of the sum of the measures. This can be done by realizing that the variance of a sum equals the sum of the variances of the individual measures plus two times the sum of all the measure covariances. Since errors of measurement are assumed to be random, they contribute to the magnitude of the measure variances but not to the magnitude of measure covariances. Hence, the greater the relative contribution of the measure covariances to the variance of the sum, the more reliable the sum. Equivalently, the more the variance of the sum exceeds the sum of the individual measure variances, the higher the sum's reliability. This rationale lies behind the derivation of Coefficient Alpha (Cronbach, 1951), estimating the reliability of a composite score made by summing individual measures:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_{SUM}^2}\right)$$

This formula is equivalent to the Spearman - Brown Prophecy Formula given above if one standardizes all variables and computes the sum of standardardized scores, thus forcing them to have equal variance. In essence, a composite score based on the simple sum of the variables weights the variables according to their variance. A sum of standardized variables weights them equally.

Although social psychologists frequently report coefficient alpha for a composite score, they are more likely to conduct a factor analysis or principal components analysis of the individual measures and then form a composite or sum of the measures weighted by their factor loadings. In essence this approach relaxes the assumptions made by the parallel forms in a different way, weighting the individual measures in computing the sum neither equally nor according to their variance but according to the magnitude of their relationship with the true score. Thus, we no longer are assuming equal true score variance in each of the measures.

In general, the principal components approach to reliability estimation involves the following steps. One takes a set of measures presumed to have a single underlying construct or factor in common. One performs a principal components analysis, with the resulting factor loadings on the first unrotated principal component estimating the correlation between each measure and the single underlying construct. The sum of these squared loadings is the component's eigenvalue or latent root, which should be large relative to the number of measures if a strong argument is to be made that the measures tap a single underlying construct. Next, a composite or weighted-sum score is computed for each individual, summing scores that are weighted by the measure's factor loading. The reliability of this composite score is given as:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{1}{\lambda}\right)$$

where λ is the eigenvalue of the first principal component. This has been shown to be the maximum possible alpha for any weighted linear combination of the component measures (Bentler, 1968). Extensions of this approach have also been worked out for the case where the set of measures are not single-factored and where multiple composite scores are derived from a rotated factor soluation (Armor, 1974).

Numeric Example. In Table 8, hypothetical data are given representing the scores of six subjects (A through F) on three alternative measures (X1, X2, and X3). Also given are 1) the means and variances of the three measures; 2) the simple sum of the three measures for each subject and its mean and variance; 3) the measure intercorrelations and covariances and 4) the loadings of the three measures on their first principal component and its eigenvalue. We have included only six subjects in these data to keep the example simple. Many more subjects would typically be needed to assess reliability and construct validity of the three measures with confidence.

Table 8

Hypothetical Data for Reliability Estimation

	SUBJECT							
	A	B	C	D	E	F	Mean	Variance
ITEM								
X1	5	3	6	5	2	2	3.833	2.967
X2	6	3	9	5	3	4	5.000	5.200
X3	9	6	7	5	5	3	5.833	4.167
Sum	20	12	22	15	10	9	14.667	28.667

Correlation (Covariance) Matrix

	X1	X2	X3
X1	1.000 (2.967)		
X2	.866 (3.400)	1.000 (5.200)	
X3	.673 (2.367)	.516 (2.400)	1.000 (4.167)

Principal Components Analysis: 1st Component Only

Eigenvalue: 2.380

	Loadings
X1	.958
X2	.902
X3	.805

The reliability of the simple sum of the measures, using the formula for coefficient alpha based on the simple sum is:

$$\alpha = \left(\frac{3}{2}\right)\left(1 - \frac{2.967 + 5.200 + 4.167}{28.667}\right) = .855$$

The average intercorrelation among the three measures equals .685. We can use the Spearman - Brown Prophecy Formula to find the reliability of the sum of the standardized measures:

$$\alpha = \frac{(3).685}{1 + (2).685} = .867$$

Equivalently, the formula for alpha can be used, forcing the three measures to have equal variance by standardizing them. Thus the variances of the standardized measures equal 1.00 and the variance of the sum of the standardized measures equals the sum of the variances plus 2 times the sum of the three intercorrelations:

$$\alpha = \left(\frac{3}{2}\right)\left(1 - \frac{1+1+1}{1+1+1+2(.866+.673+.516)}\right) = .867$$

The reliability of the optimally weighted sum, weighting the three measures according to their loadings on the first principal component is

$$\alpha = \left(\frac{3}{2}\right)\left(1 - \frac{1}{2.380}\right) = .870$$

Note that all three of these methods of weighting the three measures yield quite similar reliabilities. This will generally be the case when the variances of the measures are similar and their intercorrelations are high and relatively uniform. The three weighting methods can give radically different results, however, if the measures have substantially different variances. This is particularly likely to be the case when the measures are in different metrics (e.g., using income (in dollars) and job prestige (on a 9-point rating scale) as indicators of socioeconomic status). In such a case, the standardized or principal components weighting should be used instead of the simple sum, since measures with larger variance are weighted more heavily in the simple sum and, in the case of the use of different metrics, this would amount to weighting by a variable's metric. Additionally, when the measured intercorrelations are quite variable in magnitude, the principal components approach will yield a more reliable composite than the simple sum or standardized sum. One needs to be careful in interpreting the composite score from a principal components analysis in such a case however, since very unequal weights will be used and the composite may end up reflecting only some of the measures. When some of the intercorrelations are negative, measures need to be reverse-scored unless the principal components weighting is used.

Components of Variance and Intraclass Correlations

An alternative approach to reliability estimation partitions the variability in a subject by measure matrix of scores into its various components, following the rules of analysis of variance. The computations for the data in Table 8 yield the following values for sums of squares and mean squares due to subjects, measures, and residual or error:

Source	Sum of Squares	df	Mean Square
Between Subjects	47.778	5	9.556
Between Measures	12.111	2	6.056
Residual	13.889	10	1.389

These mean squares have expected values that are functions of the variance due to subjects, measures, and error (Cornfield & Tukey, 1956). Once these functions are specified, one can estimate these unknown variance components and from them derive reliability estimates. These reliability estimates are, in their most general form, ratios of the component of variance of interest to the sum of that component plus component(s) of error variance. They are known as intraclass correlations (Shrout & Fleiss, 1979) and we illustrate their computation in the following paragraphs.

The exact functions that relate the expected mean squares due to subjects, measures, and error to the variance components depend on whether subjects and measures are treated as random or fixed effects. In essence, a random effect involves of sampling of levels from a population of levels to which one wishes to generalize. A fixed effect is one in which all levels to which generalization is sought are included in the data collection design. We will return to this distinction in a bit and clarify its meaning for interpreting the intraclass correlations that we compute from the mean squares. We routinely treat subjects as a random effect since it is rarely the case that we wish to confine generalization to only the subjects from whom data have been collected. Whether measures should be treated as a fixed or random effect is a more difficult choice and depends on the design of anticipated future studies about which inferences are sought. Again, further discussion on this topic follows.

In general, in this two way design, the three variances components, due to subjects, measures, and error, can be estimated as follows from the mean square errors:

$$S_S^2 = \frac{MS_S - MS_E}{k}$$

$$S_M^2 = \frac{MS_M - MS_E}{n}$$

$$S_E^2 = MS_E$$

where S_S^2, S_M^2, S_E^2 are the estimated variance components associated with subjects, measures, and error, respectively, MS_S, MS_M, MS_E are the computed mean squares due to subjects, measures, and error, and k and n are the number of measures and subjects. For the data of Table 8, the estimated values of the variance components are:

$$S_S^2 = 2.722$$

$$S_M^2 = 0.778$$

$$S_E^2 = 1.389$$

If measures are considered fixed, then their variance is considered to be constant in the currrent and future studies, since in all future studies the same measures will be included. To estimate the reliability of the subjects' scores on these measures, we want to compare the component of variation due to subjects with the total variation due to subjects and error, excluding the fixed measure variance. Expressed as a function of mean squares, this ratio equals:

$$\frac{S_S^2}{S_S^2+S_E^2} = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E} = \frac{9.556 - 1.389}{9.556 + (2)1.389} = .662$$

This value is the estimated reliability of a single measure. It is called the intraclass correlation with measures fixed, ICC(3,1) in the notation of Shrout and Fleiss (1979).

Note that it is similar in value, although not identical to the average intercorrelation between the three measures computed earlier, i.e., .685. Actually it is the reliability of a single measure not assuming standardization or equal variance. One can compute it directly from the variances and covariances of the measures given in Table 8, by dividing the sum of the three covariances by the sum of the three variances:

$$\frac{3.400 + 2.367 + 2.400}{2.967 + 5.200 + 4.167} = .662$$

Thus, rather than aveage the three correlations, one pools the three covariances and the three variances and computes a single correlation from these pooled values.

The relationship between this intraclass correlation and what we earlier computed as coefficient alpha is further clarified by noting that if we use the intraclass correlation for the value of r_{ij} in the Spearman - Brown Prophecy Formula, we get a value for the reliability of the sum of the three measures that is equivalent to coefficient alpha for these three measures, allowing differences in their variances:

$$\alpha = \frac{kr_{ij}}{1 + (k-1)r_{ij}} = \frac{(3).662}{1 + (2).662} = .855$$

This is identical to the value that we computed using the formula for coefficient alpha previously.

If measures are considered random, then the measures included in the current study are only a sample of measures and different samples of measures are presumably to be used in future studies. As a result, subjects' scores may contain error in part because we employ different measures or subsets of measures with different subjects. Accordingly, measure variance should be considered a component of error variance. Hence, we want to compare the component of variation due to subjects in the current study with the total variation due to subjects, measures, and error. We can express this also as a function of the mean squares:

$$\frac{S_S^2}{S_S^2 + S_M^2 + S_E^2} = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E + k(MS_M - MS_E)/n}$$

$$= \frac{9.556 - 1.389}{9.556 + 2(1.389) + .5(4.667)} = .557$$

This is also an intraclass correlation that estimates the reliability of a single measure, ICC(2.1) in the notation of Shrout and Fleiss (1979). But this time, we are estimating the reliability of a measure allowing for the fact that different subjects may receive different measures so that variation between measures becomes a part of the error variation.

To clarify the difference between the two intraclass correlations, imagine two future studies. In the first, we use one of the present three measures for all subjects. Then the appropriate reliability to be computed from the current data is the intraclass correlation with measures fixed (i.e., .662). In the second future study, however, different measures are used for different subjects so that some of the differences from subject to subject are due to measure differences as well as to true subject differences and error. If that is the future study to which we want to generalize, then the intraclass correlation that treats measures as a random factor is the appropriate reliability estimate (i.e., .557). The two intraclass correlations will differ to the extent that the variance component due to measures is large or, equivalently, to the extent that differences among means of the measures are relatively large.

An Alternative Estimation Approach. An alternative procedure for estimating the variance components and intraclass correlation has recently been outlined by Kenny and Judd (in press). The advantage of this procedure is its flexibility. For instance, it handles randomly missing data with relative ease.

Consider the 18 scores in Table 8. The overall variance of these 18 scores can be expressed as a function of the sum of the squared differences between all possible pairs of scores. However, some of these pairs of scores involve two scores from the same subject. They thus are dependent because they come from the same subject. Other pairs are dependent because they come from the same measure. Only pairs of scores that come from different subjects and different measures are independent of each other.

In Table 9, we present a matrix of all possible pairs of the 18 scores, indicating which pairs are independent (I), which share a common subject (S), and which share a common measure (M). If observations are in fact dependent because of common subjects or common measures, then we would expect the squared differences on average to be smaller between scores that are dependent due to a common subject or a common measure than between scores that are independent. More formally, it can be shown that the expected value of the squared differences between independent pairs equals $2\sigma_I^2$, where σ_I^2 is the variance of independent scores. The expected value of the squared differences between pairs that share a common subject equals $2\sigma_I^2(1 - \rho_S)$. If pairs share the same measure, then the expected squared difference equals $2\sigma_I^2(1 - \rho_M)$. In these expressions, ρ_S and ρ_M are the correlations or dependencies in the data due to common subjects or measures. From these expressions, we can estimate the values of σ_I^2, ρ_S, and ρ_M from the computed averages of the squared differences between independent and dependent pairs of scores. Let U equal the average of the squared differences computed between pairs of scores that are independent of each other (i.e., those pairs designated I in Table 9). Let L_S equal the average squared difference between pairs that have a common subject, and let L_M equal the average squared difference between pairs that have a common measure. Then the estimates of σ_I^2, ρ_S, and ρ_M are given by the expressions:

$$S_I^2 = U/2$$

$$r_S = 1 - \frac{L_S}{U}$$

$$r_M = 1 - \frac{L_M}{U}$$

For the scores in 9, all of the pairwise squared differences are presented in Table 10. From these we can calculate the average squared difference among independent pairs, pairs with a common subject, and pairs with a common measure: U equals 9.778, L_S equals 4.333, and L_M equals 8.222. Accordingly the three estimates equal:

$$S_I^2 = 4.889$$

$$r_S = .557$$

$$r_M = .159$$

Note that the dependency due to subjects, r_S is identical to the intraclass correlation treating measures as a random factor that was computed previously. The dependency due to measures, r_M is the intraclass correlation due to measures (i.e., the ratio of the variance component due to measures to the sum of the components due to measures, subjects, and error). Additionally, the values of the variance components, due to subjects, measures, and error, can be directly calculated from the three values of U, L_S, and L_M:

$$S_S^2 = (U - L_S)/2 = (9.778 - 4.333)/2 = 2.722$$

$$S_M^2 = (U - L_M)/2 = (9.778 - 8.222)/2 = 0.778$$

$$S_E^2 = (L_S + L_M - U)/2 = (4.333 + 8.222 - 9.778)/2 = 1.389$$

The variance of the independent scores, S_I^2, equals the sum of these three components, as it should since scores that are independent from each other come from different subjects and different measures.

Table 9

Dependencies Among Pairs of Scores in Table 8

	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3
A1																		
A2	S																	
A3	S	S																
B1	M	I	I															
B2	I	M	I	S														
B3	I	I	M	S	S													
C1	M	I	I	M	I	I												
C2	I	M	I	I	M	I	S											
C3	I	I	M	I	I	M	S	S										
D1	M	I	I	M	I	I	M	I	I									
D2	I	M	I	I	M	I	I	M	I	S								
D3	I	I	M	I	I	M	I	I	M	S	S							
E1	M	I	I	M	I	I	M	I	I	M	I	I						
E2	I	M	I	I	M	I	I	M	I	I	M	I	S					
E3	I	I	M	I	I	M	I	I	M	I	I	M	S	S				
F1	M	I	I	M	I	I	M	I	I	M	I	I	M	I	I			
F2	I	M	I	I	M	I	I	M	I	I	M	I	I	M	I	S		
F3	I	I	M	I	I	M	I	I	M	I	I	M	I	I	M	S	S	

Table 10

Squared Differences Between Pairs of Scores in Table 8

	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3
A1																		
A2	1																	
A3	16	9																
B1	4	9	36															
B2	4	9	36	0														
B3	1	0	9	9	9													
C1	1	0	9	9	9	0												
C2	16	9	0	36	36	9	9											
C3	4	1	4	16	16	1	1	4										
D1	0	1	16	4	4	1	1	16	4									
D2	0	1	16	4	4	1	1	16	4	0								
D3	0	1	16	4	4	1	1	16	4	0	0							
E1	9	16	49	1	1	16	16	49	25	9	9	9						
E2	4	9	36	0	0	9	9	36	16	4	4	4	1					
E3	0	1	16	4	4	1	1	16	4	0	0	0	9	4				
F1	9	16	49	1	1	16	16	49	25	9	9	9	0	1	9			
F2	1	4	25	1	1	4	4	25	9	1	1	1	4	1	1	4		
F3	4	9	36	0	0	9	9	36	16	4	9	4	1	0	4	1	1	

In sum, this approach to estimating dependencies is identical to the variance components approach outlined above. Its advantage comes from the fact that values of U, L_S, and L_M can be computed even with substantial amounts of missing data, when either mean squares or correlations among measures could not be computed. Imagine, for instance, that each subject was missing two or even three scores on a random basis. Neither correlations between measures nor sums of squares and mean squares could be computed readily. Yet values of U, L_S, and L_M could be computed, assuming that there were some pairs in all three categories (i.e., independent, from a common subject, from a common measure). Simulations reported in Kenny and Judd (in press) reveal that the estimated intraclass correlation and variance components appear to be reasonable even with substantial missing data.

Convergent and Discriminant Validity

Almost all of what has been reviewed so far in this section of the chapter concerns the estimation of reliability, although under the variance components approach we did consider variation due to measures as well as true score variation associated with subjects and error variation. We turn now to the treatment of systematic, rather than error, variation in our measures and procedures for discriminating between systematic variation due to the construct that we wish to be measuring and systematic variation due to constructs of disinterest. We will refer to this latter variation as systematic error variation.

In general, information about convergent and discriminant validity of a particular measure is contained in the set of covariances or correlations between that measure and other measures assessing both the same and different constructs. The basic principle underlying this statement is the realization that two measures ought to covary to the extent that they measure the same or related constructs. To the extent that they measure different and unrelated constructs, their covariance should be small.

Consider two variables, X_1 and X_2. We assume that each one measures, to some extent, three different constructs, η_1, η_2, and η_3. Let us assume that η_1 represents the constuct of interest and the other two constructs are sources of systematic error. Additionally, each variable contains a certain amount of random error, ε_{1i} and ε_{2i}. For algebraic ease, we assume that all variables and constructs have expected values of zero and unit variances. Accordingly, the λ_{jk} loading coefficients indicate the extent to which latent construct k contributes to measured variable j. The following are the construct or latent variable equations that represent the components of the two variables:

$$X_{1i} = \lambda_{11}\eta_{1i} + \lambda_{12}\eta_{2i} + \lambda_{13}\eta_{3i} + \lambda_{1\varepsilon}\varepsilon_{1i}$$

$$X_{2i} = \lambda_{21}\eta_{1i} + \lambda_{22}\eta_{2i} + \lambda_{23}\eta_{3i} + \lambda_{2\varepsilon}\varepsilon_{2i}$$

Using the algebra of variances and covariances (Kenny, 1979), we can derive expectations for the variances of the two variables and for their correlation (ρ represents the expected correlation between two variables or constructs):

$$\sigma^2_{X_1} = \lambda^2_{11} + \lambda^2_{12} + \lambda^2_{13} + 2\lambda_{11}\lambda_{12}\rho_{\eta_1\eta_2} + 2\lambda_{11}\lambda_{13}\rho_{\eta_1\eta_3} + 2\lambda_{12}\lambda_{13}\rho_{\eta_2\eta_3} + \lambda^2_{1\varepsilon}$$

$$\sigma^2_{X_2} = \lambda^2_{21} + \lambda^2_{22} + \lambda^2_{23} + 2\lambda_{21}\lambda_{22}\rho_{\eta_1\eta_2} + 2\lambda_{21}\lambda_{23}\rho_{\eta_1\eta_3} + 2\lambda_{22}\lambda_{23}\rho_{\eta_2\eta_3} + \lambda^2_{2\varepsilon}$$

$$\rho_{X_1X_2} = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22} + \lambda_{13}\lambda_{23} + (\lambda_{11}\lambda_{22}+\lambda_{12}\lambda_{21})\rho_{\eta_1\eta_2} + (\lambda_{11}\lambda_{23}+\lambda_{13}\lambda_{21})\rho_{\eta_1\eta_3} + (\lambda_{12}\lambda_{23}+\lambda_{13}\lambda_{22})\rho_{\eta_2\eta_3}$$

According to these expressions, the variance of each measured variable is a function of the relative contributions of each latent construct to the variable, the correlations between the latent constructs, and random error. The correlation between the two measured variables is a complex function of the extent to which the same latent constructs contribute to both and the extent to which the latent constructs are themselves correlated. Examining the magnitude of this correlation is likely to be relatively uninformative about the construct validity of the two measured variables unless we can make certain assumptions about the make-up of the two variables. For instance, let us assume that the two constructs of disinterest are measure-specific, that is η_2 contributes only to X_1 and η_3 contributes only to X_2. In other words, we are assuming that λ_{12} and λ_{23} both equal zero. Then the expectation for the correlation between the two measured variables reduces to:

$$\rho_{X_1X_2} = \lambda_{11}\lambda_{21} + \lambda_{11}\lambda_{22}\rho_{\eta_1\eta_2} + \lambda_{13}\lambda_{21}\rho_{\eta_1\eta_3}$$

Thus, the correlation between the two measured variables is now affected solely by the extent to which they both measure the construct of interest ($\lambda_{11}\lambda_{21}$) and the extent to which the two sources of systematic error are in turn correlated with the construct of interest. Note, however, that all three terms in this equation have at least one loading coefficient of one of the measured variables on the construct of interest. As a result, the correlation between the two variables will tend to be large only if the two variables each tend to possess convergent validity. Additionally, as the correlations betwen the two sources of systematic error and the construct of interest approach zero, the equation reduces further:

$$\rho_{X_1X_2} = \lambda_{11}\lambda_{21}$$

The correlation reflects only the extent to which the two measures both measure the construct of interest. In other words, if we can make these simplifying assumptions, the correlation between two variables is indicative of the variables' joint convergent validity.

These assumptions are, of course, highly restrictive. Nevertheless, they provide the justification for the reliance upon the "multiple operations" approach to construct validity (Campbell, 1960; Cambpell & Fiske, 1959; Cook & Campbell, 1979; Cronbach & Meehl, 1955; Judd & Kenny, 1981; Webb, Campbell, Schwartz, & Sechrest, 1966). Multiple measures of the same construct should be highly related to each other if they in fact do jointly measure the construct of interest. But high intercorrelations are not a sufficient condition for claiming convergent validity. Additionally, one must have confidence that the multiple measures do not share other systematic sources of variation or constructs of disinterest. In other words, one needs multiple measures that are as dissimilar as possible in terms of their irrelevancies or sources of systematic error.

Although most social psychologists have learned the lesson that multiple operations are important, all too often we are content to use multiple measures whose irrelevancies may be highly redundant. Thus, we tend to use redundant questionnaire items and convince ourselves of their validity by showing that they are highly intercorrelated (i.e., a high coefficient alpha). But high internal consistency is not the same thing as convergent validity. We need additionally to be convinced that they do not share large components of systematic error.

Discriminant validity is demonstrated by showing that measures of the construct of interest intercorrelate more highly with each other than they do with measures that assess the construct of disinterest against which discriminant validity is sought. In its most stringent form, one would like evidence that measures of the construct of interest are uncorrelated with measures of the construct of disinterest. But such a situation should be expected only if the two constructs are themselves unrelated to each other. A more reasonable expectation is that measures of the construct of interest should be more highly correlated with each other than they are with measures of the construct of disinterest.

Multitrait - Multimethod Correlation Matrix. Evidence for simultaneous convergent and discriminant validity derives from the multitrait-multimethod correlation matrix, as outlined by Campell and Fiske (1959). Such a matrix involves multiple measures of two or more constructs, crossing alternative methods of measurement with constructs. For instance, a social psychologist interested in person perception might collect data measuring the sociability, creativity, and intelligence of a set of target individuals. Each of the two traits is measured in multiple ways: self-ratings provided by the target, ratings of the target provided by a close friend, and performance scores on inventories designed to measure the three constructs. The hypothetical correlation matrix of Table 11 results. Scores on each measure are assumed to be a function of three things: the relevant trait construct (sociable vs. creative vs. intelligent), systematic error due to method (self-rating vs. friend rating vs. inventory), and random error. Following the logic summarized above when we derived the factors contributing to the correlation between two measured variables, the correlations in the present matrix are due to:

1) convergent validity between measures of the same trait using different methods;

2) shared systematic error variance due to measures using the same method;

3) correlations between the latent trait constructs; and

4) correlations between the latent method constructs.

In addition to providing evidence for convergent validity, the matrix provides evidence for two sorts of discriminant validity. First, it enables the researcher to examine the extent to which systematic error variance due to method contributes to the observed measures. Thus, it discriminates between true-score trait variance and method variance. Additionally, it permits the researcher to examine whether the latent trait constructs are discriminable.

Table 11

Multitrait - Multimethod Matrix: 3 Traits by 3 Methods

Method	Trait	S - S	S - C	S - I	F - S	F - C	F - I	I - S	I - C	I - I
Self	Sociable									
	Creative	*.34*								
	Intelligent	*.27*	*.45*							
Friend	Sociable	**.58**	.18	.20						
	Creative	.14	**.42**	.09	*.29*					
	Intelligent	.19	.12	**.46**	*.21*	*.38*				
Inventory	Sociable	**.53**	.14	.05	**.60**	.12	.03			
	Creative	.06	**.31**	.12	.10	**.34**	.09	*.18*		
	Intelligent	.12	.14	**.47**	.08	.15	**.43**	*.14*	*.24*	

Campbell and Fiske (1959) outlined an informal logic for the analysis of multitrait - multimethod matrices. This logic centers around the values of three sorts of correlations in the matrix:

1) Correlations between measures of the same traits using different methods. Campbell and Fiske (1959) called these validity correlations. They are in bold type in Table 11.

2) Correlations between different traits using the same methods. These are italicized in Table 11.

3) Correlations involving measures of different traits with different methods.

Campbell and Fiske (1959) argued that four criteria should be met if the correlation matrix showed evidence of convergent and discriminant validity:

1) correlations involving measures of the same trait with different methods (i.e., the validity correlations) should be substantially larger than zero.
2) correlations involving measures of the same trait with different methods should be larger than correlations involving measures of different traits with different methods.
3) correlations involving the same trait but different methods should be higher than correlations involving different traits and the same method.
4) Roughly the same pattern of correlations should be observed between measures of the different traits within each of the types of methods.

The first two criteria suggest convergent validity. The third argues that the contribution of trait construct variance to the measures is larger than method variance, thus suggesting discriminant validity against method constructs. And the fourth criteria suggests that the underlying trait constructs are consistently discriminable regardless of method.

At a later point in this section of the chapter, we will examine more formal procedures for analyzing multitrait-multimethod matrices. These analyses permit the researcher to estimate the contribution of trait variance, method variance, and error variance to each of the measures, as well as to estimate the correlations between the trait constructs and between the method constructs. In spite of these analytic advances, however, the informal criteria outlined by Campbell and Fiske (1959) continue to have great intuitive appeal.

Variance Components and Generalizability Theory. In our earlier discussion of reliability that focused on the estimation of intraclass correlations, we raised the possibility that variation due to measures might be considered to be an aspect of error variation depending on the type of generalization which one sought to make. If, for instance, in an anticipated study, different individuals were to receive different measures, then differences between measures would constitute a component of unreliability and accordingly ought to be included in the denominator of the intraclass correlation. An alternative way to understand this result is to argue that an individual's score is in part due to his or her true standing on the construct of interest, in part due to the systematic error associated with a particular measure, and in part random error. Thus, the earlier analysis that separated random error from systematic error due to measure in fact was exploring issues of convergent and discriminant validation.

This approach has been importantly extended by Lee Cronbach and others into what has become widely known as Generalizability Theory (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Fyans, 1983). The fundamental insight of this theory is that any score has potentially many factors that contribute to it: the person being measured, the measure being used, the experimenter administering the measure, the setting, the season of the year, the time of the day, etc. And all of these factors (or facets in the terms of Generalizability Theory) affect the magnitude of error in the score and the degree to which one can generalize from it. Thus, if generalization across measures is wanted, then variation due to measures needs to be included as a component of error. Similarly, if one wishes to generalize across experimenters or settings, then these too need to be considered as components of systematic error. The theory therefore takes issue with the assumption of classic test theory that there is a single error component. Rather, there are multiple facets that contribute to each score and whether or not these are considered components of error depends on the situation to which one wishes to generalize. Accordingly, there is no single index of reliability; rather there are alternative generalizability coefficients dependent on the generalization intended.

Analytically, the approach of Generalizability Theory is an experimental one in which variance components associated with multiple facets are estimated from the mean squares that one calculates through analysis of variance, just as we did above in the case of the design where subjects were crossed with measures. This earlier design is considered to be a one-facet design, since facets are defined as systematic variation in components other than subjects. A two facet design would be one in which both measures and occasions varied, and so forth. The estimation of the relevant components of variance depends on whether facets are crossed with each other or nested and on whether their levels are considered to be fixed or random. As always, subjects are considered to be a random effect and are generally, although not always, crossed with facets. Most typically, facets are considered to be random, since we typically wish to generalize to situations involving other measures, occasions, experimenters, and so forth. The important point, however, is that the definition of the facets, whether they are considered to be random or fixed, and whether they are considered to be a component of error in calculating generalizability coefficients (or intraclass correlations) depends on the situation to which one would like to generalize.

General rules for deriving expected values for mean squares given a wide variety of designs are provided in books devoted to analysis of variance and experimental design (e.g., Kirk, 1982; Winer, Brown, & Michels, 1991). These generally follow the classic exposition on the subject set out by Cornfield and Tukey (1956). The expected mean squares are functions of the components of variance attributable to facets, subjects, and their interactions. One can derive estimates of the components by substituting the obtained mean squares for their expected values. We followed that procedure earlier in considering the design in which subjects were crossed with measures. We illustrate it now with a two facet design.

Suppose we wished to measure subjects' tendency to display prejudice towards outgroup members. We sample 100 majority group subjects and ask them to indicate their agreement with three alternative measures of outgroup derogation, directed towards three different outgroups. The three measures are:

1) "_______ are too pushy in insisting on their rights in our society."

2) "Affirmative action really isn't needed anymore to guarantee the fair treatment of ________ in our society."

3) "_______ don't appreciate the extent to which you have to work hard to make it in our society."

These items are given to subjects, referencing three different minority outgroups, substituting "Blacks", "Hispanics", and "Native Americans" in turn for the blanks in each item. In total then, each subject gives nine ratings.

From the resulting data, we estimate the sums of squares and mean squares due to subjects, measures, outgroups, and their interactions. These values are given at the top of Table 12. Treating all three factors as random, the expected values of these mean squares are given in the middle panel of Table 12. Here n_S is the number of subjects and σ_S^2 is the variance component due to subjects. Other subscripts reference measures (M) and groups (G). Variance components for interactions are indicated by products of the subscripts. The triple interaction is confounded in this design with residual error. Its variance component has an R subscript. From the expressions for the expected values of the mean squares and their actual computed value from the data, one can derive estimates of the variance components. These estimates are given at the bottom of Table 12.

Table 12

Variance Components Analysis of a Two Facet Randomized Design

Source Table from Data

Source	Sum of Squares	df	Mean Square
Subjects	12525.92	99	126.52
Measures	28.86	2	14.43
Groups	722.41	2	361.21
Subjects X Measures	2540.10	198	12.83
Subjects X Groups	10259.35	198	51.81
Measures X Groups	37.33	4	9.33
Residual	3305.65	396	8.35

Expected Mean Squares

$$EMS_S = n_M n_G \sigma^2_S + n_M \sigma^2_{SG} + n_G \sigma^2_{SM} + \sigma^2_R$$

$$EMS_M = n_S n_G \sigma^2_M + n_S \sigma^2_{MG} + n_G \sigma^2_{SM} + \sigma^2_R$$

$$EMS_G = n_S n_M \sigma^2_G + n_S \sigma^2_{MG} + n_M \sigma^2_{SG} + \sigma^2_R$$

$$EMS_{SM} = n_G \sigma^2_{SM} + \sigma^2_R$$

$$EMS_{SG} = n_M \sigma^2_{SG} + \sigma^2_R$$

$$EMS_{MG} = n_S \sigma^2_{MG} + \sigma^2_R$$

$$EMS_R = \sigma^2_R$$

Estimated Variance Components

$S^2_S = 7.80$

$S^2_M = .00$

$S^2_G = 1.03$

$S^2_{SM} = 1.49$

$S^2_{SG} = 14.49$

$S^2_{MG} = .01$

$S^2_R = 8.35$

The estimated variance components themselves are readily interpretable. The variance component due to subjects represents the variation between subjects in their habitual or average response regardless of the outgroup being rated or the measure used. In the present context, it can be interpreted as individual variation in generalized prejudice, regardless of the outgroup and the question asked. Similar interpretations can be made for the variance component due to groups or measures. For instance, the variance component due to groups tells us about variation between groups in prejudice expressed towards them, collapsing across methods and subjects.

The components associated with the subject by group and subject by measures interactions are also of interest. The subject by group interaction estimates the extent to which subjects respond differently to the different outgroups. Higher values would indicate that individual differences in prejudice depend on the outgroup towards whom prejudice is expressed. The subject by measure interaction represents variation in expressed prejudice as a function of the measure used.

A variety of different generalizability coefficients (which are generalizations of the intraclass correlations computed earlier) can be computed from these variance components, depending on the nature of future research to which one would like to generalize. In general, the form of these coefficients is the ratio of the variance component(s) of interest to the variance of the observed score in the anticipated study. Thus, for instance, suppose we wanted to know the reliability of an individual's score in a future study where different subjects would be asked about different groups. The appropriate coefficient would then tell us about the ratio of the variance component due to subject to the sum of the variance components due to subject, group, subject by group, and residual:

$$\frac{S_S^2}{S_S^2 + S_G^2 + S_{SG}^2 + S_R^2} = \frac{7.80}{7.80 + 1.03 + 14.49 + 8.35} = .25$$

As is apparent from this calculation, one reason why this coefficient is not very large is because there is substantial variation associated with the subject by group interaction (S^2_{SG}). The relative magnitude of this component tells us that expressed prejudices are relatively group specific. Equivalently, there seems to be a fair amount of discriminant validity among the prejudices towards individual groups. Hence, in an anticipated study where different subjects are asked about different groups, the reliability of responses suffers because group varies between subjects. On the other hand, if in a future study group was held constant, then we would expect greater reliability.

This discussion illustrates how the variance components inform us about convergent and discriminant validity in this two-facet design as well as about anticipated reliability. If prejudices are group specific, that is if there exists discriminant validity between prejudice expressed towards one outgroup and that expressed towards another, then we should find the variance component due to the subject by group interaction to be large relative to the variance component due to subjects. Kenny (1994) shows that the average discriminant validity correlation between prejudices expressed towards different outgroups in this design equals

$$\frac{S^2_S}{S^2_S + S^2_{SG}} = \frac{7.80}{7.80 + 14.49} = .35$$

In other words, over and above measure differences, the correlation between expressed prejudice toward one outgroup in these data and another equals on average only .35.

These components of variance tell us not only about discriminant validity between prejudice directed towards one outgroup versus another but also about whether different measures give discriminably different results. In this case, of course, we would like to argue against discriminant validity due to measures, that is, we would like to argue that one gets the same answers regardless of which measure is used. The parallel discriminant validity correlation for measures is

$$\frac{S^2_S}{S^2_S + S^2_{SM}} = \frac{7.80}{7.80 + 1.49} = .84$$

Thus, these data and their associated variance components suggest that although there are stable individual differences in prejudice regardless of the outgroup against which prejudice is expressed, there is also substantial evidence for discriminant validity in the expressions of prejudice towards one outgroup as opposed to another. Additionally, these data suggest that method variance is not substantial and that expressions of prejudice from one measure to another are hardly discriminable.

It probably has occurred to the reader by now that the data we have been using here to compute variance components and illustrate their utility are in form no different from the data that are used to construct a multitrait - multimethod matrix. If we assume that different outgroups constitute different traits and different methods are different measures, then the correspondence is exact. In Table 13 we present the multitrait - multimethod correlation matrix computed from the same raw data that were used to conduct the analysis of variance reported in Table 12. Not only are these data parallel in form to data used to construct a multitrait - multimethod matrix, but also we have used the variance component approach to examine the issues of construct validity typically assessed by examining a multitrait - multimethod correlation matrix. Specifically, we have examined whether these data exhibit discriminant validity between traits (i.e., between outgroups) and if there is substantial variation due to methods. Later in this section we will illustrate a more exact parallel between the analysis of variance or variance component approach to the multitrait - multimethod matrix and that based on examining the resulting correlations or covariances (following the lead of Kenny, 1994, also discussed in Cronbach et al., 1972).

Table 13

Means, Variances, and Correlations (Covariances) of Subjects by Groups
by Measures Data Reported in Table 12

(Variables Defined as Ratings of Specific Groups on Specific Measures;
Correlations Computed across Subjects;
Means (Variances) Reported in Diagonal Cells)

Group	Measure	G1M1	G1M2	G1M3	G2M1	G2M2	G2M3	G3M1	G3M2	G3M3
1	1	8.64								
		(24.40)								
1	2	.66	9.37							
		(18.05)	(30.72)							
1	3	.60	.71	8.81						
		(15.77)	(20.78)	(28.17)						
2	1	.35	.34	.22	9.63					
		(10.88)	(11.92)	(7.45)	(40.29)					
2	2	.31	.37	.26	.75	9.48				
		(9.34)	(12.77)	(8.55)	(29.38)	(38.15)				
2	3	.16	.18	.14	.69	.72	10.01			
		(4.62)	(5.62)	(4.42)	(25.12)	(25.53)	(33.27)			
3	1	.26	.24	.15	.44	.34	.22	10.72		
		(7.06)	(7.53)	(4.56)	(15.62)	(11.62)	(6.99)	(31.08)		
3	2	.12	.20	.19	.34	.27	.12	.69	11.26	
		(3.46)	(6.39)	(5.74)	(12.33)	(9.59)	(3.91)	(22.28)	(33.11)	
3	3	.19	.30	.26	.32	.33	.30	.68	.73	11.33
		(5.22)	(9.19)	(7.45)	(11.22)	(11.19)	9.49)	(20.62)	(23.09)	(30.02)

We have only illustrated the variance components approach and Generalizability Theory in a relatively simple case. The theory has been extended to handle a variety of designs more complex than the one we have used as an illustration. Additionally, the theory takes into account the sort of decision to be made about subjects in future studies, for instance whether one is solely interested in identifying the relative position of each subject compared to the others or whether one wishes to ascertain the absolute position of subjects on a measurement scale. Finally, the theory has been extended to the multivariate case. The point of our exposition is simply to familiarize social psychologists with this approach to systematic error. It seems to be potentially a very fruitful approach that is at present underutilized in the social psychological literature. While one encounters studies that estimate the reliability of judges or measures by reporting an intraclass correlation, following the procedures discussed earlier in this section, seldom does one encounter the sort of systematic variation of multiple facets that permits the generalization of measurement results outlined by Cronbach and his colleagues. We believe that more careful attention to multiple facets and their role in inducing systematic error variation is appropriate in social psychology. Additionally, we think the computation of estimated variance components and their comparison has much to recommend it.

Construct Validation Through Confirmatory Factor Analysis

The distinct advantage of the variance components approach is that through experimental manipulation of multiple facets, one can directly estimate the contribution of various sources of systematic error to the measured variable(s). One would like to be able to do this also in situations where the components of error have not been manipulated systematically. Let us revisit the hypothetical construct equations for two variables, X_1 and X_2, presented earlier in this section:

$$X_{1i} = \lambda_{11}\eta_{1i} + \lambda_{12}\eta_{2i} + \lambda_{13}\eta_{3i} + \lambda_{1\varepsilon}\varepsilon_{1i}$$

$$X_{2i} = \lambda_{21}\eta_{1i} + \lambda_{22}\eta_{2i} + \lambda_{23}\eta_{3i} + \lambda_{2\varepsilon}\varepsilon_{2i}$$

It would be useful to estimate the loading coefficients in these construct equations, the λ_{jk}'s, as well as the correlations among the latent constructs, η_k even if those latent constructs have not been systematically varied through facet manipulations. Direct estimation of these unknowns would provide maximal information relevant to the variables' construct validity.

Depending on the nature of the hypothesized latent construct model underlying each measured variable and on the number of measured variables, it is sometimes possible to accomplish this direct estimation through the use of confirmatory factor analysis.

Confirmatory factor analysis is a general procedure for estimating the parameters of a certain class of structural equations models (Bentler, 1979; Jöreskog, 1981; Judd, Jessor, & Donovan, 1986; Kenny, 1979; Loehlin, 1992) in which one wishes to estimate the loading coefficents of observed variables on latent or hypothetical constructs and the variances and covariances of these latent constructs. Structural parameters linking the latent constructs to each other are not hypothesized in confirmatory factor analysis models although they are estimable in the more general class of structural equation models. As we will see, in addition to estimating the parameters of construct equations (the loadings and construct variances and covariances), confirmatory factor analysis can also in some cases test the consistency of the hypothezied construct model with the observed data. Thus, like the representational approaches to measurement, the use of confirmatory factor analysis in evaluating construct validity provides the possibility of testing a hypothesized measurement model.

To provide an accessible introduction to confirmatory factor analysis and its utility in assessing construct validity, we initially will make the assumption that all variables and constructs are standardized (expectations of zero and unit variances). We will later relax this assumption in the application of confirmatory factor analysis to particular models.

Consider a very simple construct model in which two observed variables (or indicators in the language commonly used in confirmatory factor analysis) are hypothesized to derive jointly from a single latent construct, with their residual variances being attributable to random error:

$$X_{1i} = \lambda_1 \eta_i + \lambda_{1\varepsilon} \varepsilon_{1i}$$

$$X_{2i} = \lambda_2 \eta_i + \lambda_{2\varepsilon} \varepsilon_{2i}$$

The expected value of a product of any two standardized variables is their correlation. Accordingly, the correlation between these two variables can be expressed as the expected value of the product of their two construct models:

$$r_{X_{1i}X_{2i}} = \mathbf{E}\{X_{1i}X_{2i}\} = \mathbf{E}\{(\lambda_1 \eta_i + \lambda_{1\varepsilon} \varepsilon_{1i})(\lambda_2 \eta_i + \lambda_{2\varepsilon} \varepsilon_{2i})\}$$

Since the errors in these variables are assumed to be random, the expected value of products involving ε_{1i} and ε_{2i} are all zero. Hence, this expression reduces to:

$$r_{X_{1i}X_{2i}} = \lambda_1 \lambda_2 \mathbf{E}\{\eta_i^2\} = \lambda_1 \lambda_2$$

If we make the assumption that these two variables are equally valid indicators (i.e., $\lambda_1 = \lambda_2$), then their loading coefficients on the latent construct equal the square root of their correlation:

$$\lambda_1 = \lambda_2 = \sqrt{r_{X_{1i}X_{2i}}}$$

Note that this solution for the loading coefficients is equivalent to the estimation of the reliability of two variables, assuming that they are each "parallel forms," according to the classic test theory derivation given at the start of this section of the chapter.

The procedure that we have just illustrated for this very simple two variable, single construct model can be generalized to much more complex models. In essence, the expected value of the product of two observed standardized variables equals their correlation and this correlation can be expressed as the expected value of the product of their two construct equations. One can then simplify this latter expected value so that the correlation between the observed variables equals a function of the loading coefficients (i.e., the λ's) and correlations between the latent constructs (i.e., the η's). One does this for all pairs of observed variables, with the result being a set of simultaneous equations in which the observed correlations are expressed as functions of the unknown parameters (i.e., loading coefficients and latent variable correlations). Given that the number of unknown parameters is less than or equal to the number of correlations between observed variables, one can then solve for the unknown loading coefficients and latent construct correlations. We illustrate this general approach with a few more simple examples borrowed from Kenny (1979) and other classic sources.

Suppose we had three indicators of a single latent construct. Thus, our three construct equations are:

$$X_{1i} = \lambda_1\eta_i + \lambda_{1\varepsilon}\varepsilon_{1i}$$
$$X_{2i} = \lambda_2\eta_i + \lambda_{2\varepsilon}\varepsilon_{2i}$$
$$X_{3i} = \lambda_3\eta_i + \lambda_{3\varepsilon}\varepsilon_{3i}$$

We now have three correlations among observed variables and these are functions of the unknown loading coefficients:

$$r_{X_{1i}X_{2i}} = \lambda_1\lambda_2$$
$$r_{X_{1i}X_{3i}} = \lambda_1\lambda_3$$
$$r_{X_{2i}X_{3i}} = \lambda_2\lambda_3$$

Unlike the previous case with only two observed variables, we now have as many unknown parameters to be estimated as observed correlations. Hence we have an exact solution for the three unknown parameters without making the equality assumption we did in the two indicator case. The estimates of the unknown loadings are:

$$\lambda_1 = \sqrt{\frac{r_{X_{1i}X_{2i}}\, r_{X_{1i}X_{3i}}}{r_{X_{2i}X_{3i}}}}$$

$$\lambda_2 = \sqrt{\frac{r_{X_{1i}X_{2i}}\, r_{X_{2i}X_{3i}}}{r_{X_{1i}X_{3i}}}}$$

$$\lambda_3 = \sqrt{\frac{r_{X_{1i}X_{3i}}\, r_{X_{2i}X_{3i}}}{r_{X_{1i}X_{2i}}}}$$

Adding a fourth indicator to this single latent factor model gives us six equations with only four unknown loading coefficients:

$$r_{X_{1i}X_{2i}} = \lambda_1\lambda_2$$
$$r_{X_{1i}X_{3i}} = \lambda_1\lambda_3$$

$$r_{X_{1i}X_{4i}} = \lambda_1\lambda_4$$

$$r_{X_{2i}X_{3i}} = \lambda_2\lambda_3$$

$$r_{X_{2i}X_{4i}} = \lambda_2\lambda_4$$

$$r_{X_{3i}X_{4i}} = \lambda_3\lambda_4$$

These yield three solutions for each of the unknown parameter estimates. For instance,

$$\lambda_1 = \sqrt{\frac{r_{X_{1i}X_{2i}}r_{X_{1i}X_{3i}}}{r_{X_{2i}X_{3i}}}} = \sqrt{\frac{r_{X_{1i}X_{2i}}r_{X_{1i}X_{4i}}}{r_{X_{2i}X_{4i}}}} = \sqrt{\frac{r_{X_{1i}X_{3i}}r_{X_{1i}X_{4i}}}{r_{X_{3i}X_{4i}}}}$$

As a final illustration, consider a model in which two indicators load on one latent construct, two load on a second, and the two latent constructs are correlated:

$$X_{1i} = \lambda_{11}\eta_{1i} + \lambda_{1\varepsilon}\varepsilon_{1i}$$

$$X_{2i} = \lambda_{21}\eta_{1i} + \lambda_{2\varepsilon}\varepsilon_{2i}$$

$$X_{3i} = \lambda_{32}\eta_{2i} + \lambda_{3\varepsilon}\varepsilon_{3i}$$

$$X_{4i} = \lambda_{42}\eta_{2i} + \lambda_{4\varepsilon}\varepsilon_{4i}$$

The expressions for the six observed correlations are:

$$r_{X_{1i}X_{2i}} = \lambda_{11}\lambda_{21}$$

$$r_{X_{1i}X_{3i}} = \lambda_{11}\lambda_{32}\rho_{\eta_{1i}\eta_{2i}}$$

$$r_{X_{1i}X_{4i}} = \lambda_{11}\lambda_{42}\rho_{\eta_{1i}\eta_{2i}}$$

$$r_{X_{2i}X_{3i}} = \lambda_{21}\lambda_{32}\rho_{\eta_{1i}\eta_{2i}}$$

$$r_{X_{2i}X_{4i}} = \lambda_{21}\lambda_{42}\rho_{\eta_{1i}\eta_{2i}}$$

$$r_{X_{3i}X_{4i}} = \lambda_{32}\lambda_{42}$$

In these equations, $\rho_{\eta_{1i}\eta_{2i}}$ is the unknown correlation between the two latent constructs.

Now we have five unknowns in six equations (one for each observed correlation) and again the solution is overdetermined. Each loading coefficient has two solutions, for instance:

$$\lambda_{11} = \sqrt{\frac{r_{X_{1i}X_{2i}} r_{X_{1i}X_{3i}}}{r_{X_{2i}X_{3i}}}} = \sqrt{\frac{r_{X_{1i}X_{2i}} r_{X_{1i}X_{4i}}}{r_{X_{2i}X_{4i}}}}$$

and the latent construct correlation also has two solutions:

$$\rho_{\eta_{1i}\eta_{2i}} = \sqrt{\frac{r_{X_{1i}X_{3i}} r_{X_{2i}X_{4i}}}{r_{X_{1i}X_{2i}} r_{X_{3i}X_{4i}}}} = \sqrt{\frac{r_{X_{1i}X_{4i}} r_{X_{2i}X_{3i}}}{r_{X_{1i}X_{2i}} r_{X_{3i}X_{4i}}}}$$

In the first case we examined, with a single correlation between two variables and two loading parameters of those variables on their single latent construct, we had insufficient information to solve for the loading coefficients unless we assumed that they were equivalent. In the last two examples, first with four indicators of a single latent construct and then with two indicator each of two correlated constructs, we had more equations than we needed to derive the unknown parameter estimates. The issue of whether there exists a solution for the unknown parameters or whether in fact multiple solutions are possible constitutes the issue of model identification. Models are underidentified when the number of unknown parameters to be estimated exceeds the number of known variances and covariances (correlations in the standardized case) and, as a result, no solution of the simultaneous equations is possible. Models in which the number of unknowns exactly equals the number of equations are said to be just identified. Models such as the last two we presented, with more equations than unknowns and hence multiple solutions, are overidentified.

The issue of identification is crucial, for only if a model is identified can we estimate the loading coefficients and construct correlations that can be used to examine the construct validity of measured variables. Kenny, Kashy, and Bolger (in press) give a set of rules for determining whether a model is identified. In general, the simpler the construct theory or model that is hypothesized to underlie the measured variables, the greater the probability that the model will be identified. But one should certainly not posit an unrealistically simple model just to achieve identification. One needs to build an adequate measurement model for the variables at hand and then explore issues of identification. If the model is not identified, then constraints need to be placed on further data collection that permit identification. In fact, this is what generalizability theory is doing by orthogonally manipulating multiple facets or constructs. By doing this, the theory assures that the latent constructs are uncorrelated with each other, thus reducing the number of unknown parameters to be estimated.

When a model is overidentified, the presence of multiple simultaneous solutions provides a mechanism for examining the consistency of the hypothesized construct model with the data from which the model's parameters are estimated. For if the model is appropriate for the data at hand, then the multiple parameter solutions ought to give equivalent answers, within the limits of sampling error. In other words, the model implies that the alternative solutions for the unknown parameters should be equal to each other. If they are not, then we have evidence that the hypothesized construct model is inappropriate for the data at hand. In this sense, confirmatory factor analysis can provide a test of the adequacy of a construct theory, or a set of construct validity equations, given overidentification.

The general formulation for confirmatory factor analysis posits that observed variables are functions of latent constructs with unknown parameters of loading coefficients and construct variances and covariances (correlations in the standardized case). The construct model implies that the variances and covariances (again, correlations in the standardized case) among the observed variables can be expressed as functions of the unknown model parameters. Given identification, one then estimates the unknowns by solving the system of simultaneous equations. An algebraically efficient approach to this solution, given a large number of such equations and overidentification (where multiple solutions exist) is an iterative approach in which arbitrary initial values of the unknown parameters are used to generate a predicted variance / covariance matrix among the observed variables and then these parameter values are adjusted across iterations such that the discrepancies between the predicted variance / covariance matrix and the one observed in the data is minimized. Typicially a maximum likelihood discrepancy function is used. Software that accomplishes this estimation includes the LISREL program (Jöreskog & Sörbom, 1993), EQS (Bentler, 1993), and PROC CALIS in SAS (SAS Institute, 1990).

Given overidentification and the presence of sampling error, there will never be perfect convergence between the predicted matrix and the observed data matrix. Assuming multivariate normality, a maximum likelihood solution, and a sufficiently large sample, a function of the discrepancy is distributed as a chi-square statistic. This statistic can be used to test whether the predicted and observed variance / covariance matrices reliably differ from each other. If they do, then the model is revealed to be inconsistent with the data, since no solution exists that satisfies the multiple constraints resulting from overidentification. If the resulting chi-square is not significant, then the model is said to be consistent with the observed sample data.

This test of model consistency is unlike most traditional statistical inference tests, in that we wish to argue that a hypothesized construct model is consistent with the sample data, thereby wanting to verify the null hypothesis that discrepancies between the observed and predicted variance / covariance matrices are not reliable. As a result, all we can conclude is that the model and the data are not inconsistent with each other. The model can never be verified since alternative construct theories or models might be equally or more consistent with the sample data. Additionally, considerations of statistical power run counter to normal notions about the value of large samples. Although adequate sample size is necessary to meet the assumption that the maximum likelihood statistic is in fact distributed as chi-square, an exceedingly large sample will lead to the rejection of all models since substantively trivial discrepancies between the predicted and observed variance - covariance matrices will emerge as reliable. A variety of other goodness-of-fit statistics have been proposed to evaluate model consistency, given simulation results showing that the chi-square statistic is overly conservative, leading to inappropriate model rejections (Bentler, 1990; Bollen & Long, 1993, Jaccard & Wan, 1995). Many of these focus on the comparative fit of a given model with some alternative model, usually one in which the observed variables are assumed to be independent. Bentler's Comparative Fit Index (CFI; Bentler, 1990), which varies between 0 and 1, with reasonable models having CFI's of .90 or greater, seems to perform well in a variety of circumstances.

The final point to be made before illustrating the use of confirmatory factor analysis is that one can compare nested models to each other. Imagine that we had a model with four indicators all loading on a single latent construct. Initially we would estimate this model allowing the loadings to vary across the indicators. We then might want to compare this model to one in which the four indicators are assumed to have equal loadings The second model is said to be nested under the first since relaxing the constraint of equality of loadings makes it the first model. We can test whether the additional constraints lead to a significant deterioration in the fit of the model by subtracting the chi-square of the less constrained model from that of the more constrained model. The difference is itself distributed as a chi-square, with degrees of freedom equal to the difference in the two original degrees of freedom. A significant chi-square indicates that the more constrained model is reliably less consistent with the sample data that the less constrained one.

Example 1: Measures of Attitudes Toward Deviance. Judd, Jessor, and Donovan (1986) used confirmatory factor analysis to explore the construct validity of nine measures designed to assess adolescents' attitudes toward deviant behaviors. We recapitulate a portion of their analysis here as our initial illustration.

The data were collected as part of a larger longitudinal study of social development from adolescence through young adulthood. Descriptions of the full sample and methods can be found in Jessor and Jessor (1977) and Jessor, Donovan, and Costa (1991). The data reported by Judd, Jessor, and Donovan (1986) and used here were gathered from 153 seventh graders who were asked to indicate how "wrong" various deviant behaviors were. The exact questions are given in Table 14 along with the resulting correlation matrix.

Table 14

Attitude Towards Deviance Items from Judd, Jessor, and Donovan (1986)

<u>ATD Items</u> (Scale 0 = not wrong at all; 9 = very wrong)

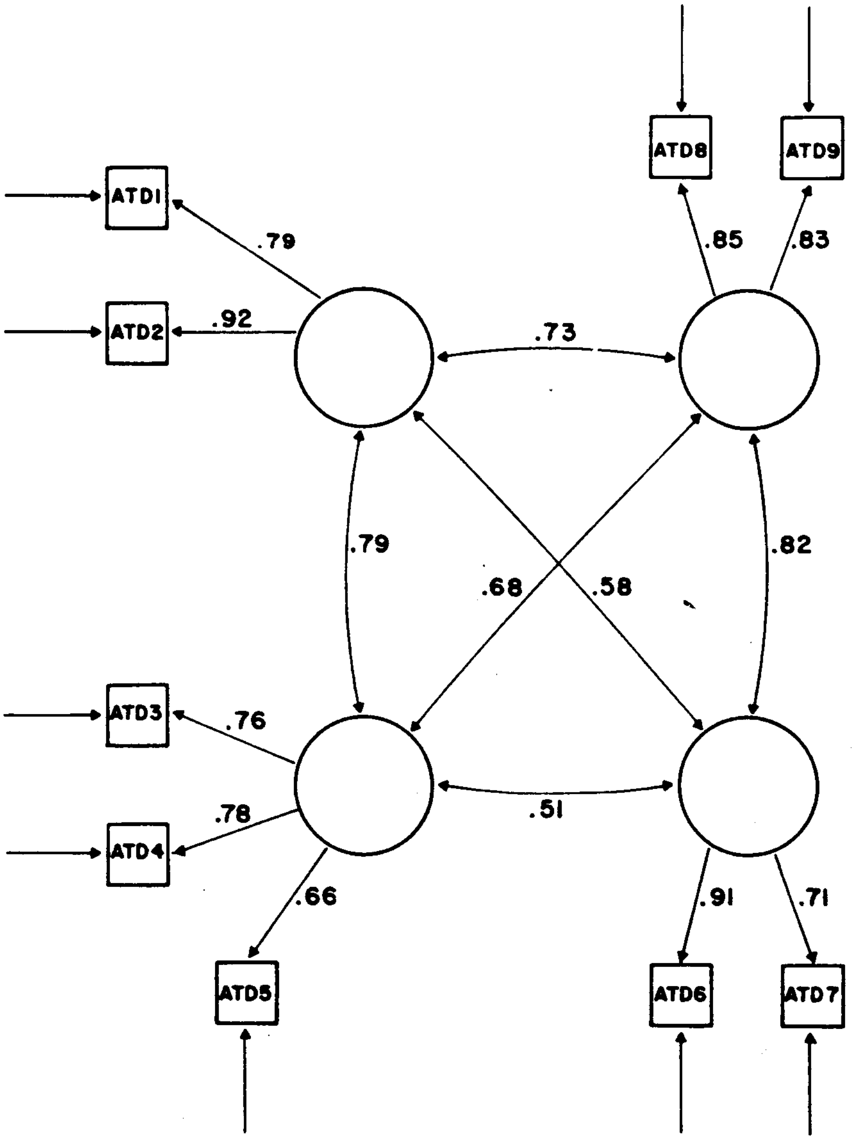
1. How wrong is it to take little things that don't belong to you?
2. How wrong is it to take something from a store without paying for it?
3. How wrong is it to give your teacher a fake excuse for missing an exam or being absent?
4. How wrong is it to lie to your parents about where you have been or who you were with?
5. How wrong is it to lie about your age when applying for a license or a job?
6. How wrong is it to beat up another kid without much reason?
7. How wrong is it to get into fist fights with kids?
8. How wrong is it to damage public or private property that does not belong to you just for fun?
9. How wrong is it to damage school property on purpose - like library books, or musical instruyments, or gym equipment?

<u>Correlation Matix</u>

	ATD1	ATD2	ATD3	ATD4	ATD5	ATD6	ATD7	ATD8	ATD9
ATD1	1.00	.	.	.	.	.	.	.	.
ATD2	.73	1.00	.	.	.	.	.	.	.
ATD3	.54	.52	1.00	.	.	.	.	.	.
ATD4	.55	.57	.62	1.00	.	.	.	.	.
ATD5	.43	.52	.54	.53	1.00	.	.	.	.
ATD6	.32	.43	.28	.38	.38	1.00	.	.	.
ATD7	.31	.41	.26	.36	.35	.67	1.00	.	.
ATD8	.42	.59	.31	.44	.35	.61	.46	1.00	.
ATD9	.43	.54	.43	.41	.38	.61	.51	.69	1.00

Problem-behavior theory (Jessor & Jessor, 1977) predicts that there ought to be a general attitude toward deviant behavior and that all of these items should tap this single underlying construct. Accordingly, the construct theory is that each variable should load on a single latent construct with uncorrelated residual variation. The parameters of this model were estimated with SAS's PROC CALIS using a maximum likelihood criterion.[4] The resulting standardized loadings of the measures on the single latent construct varied between .57 and .80 and are reported in Figure 1 of Judd, Jessor, and Donovan (1986). The magnitude of these loadings (which are estimates of the item correlations with the latent construct) indicates considerable convergent validity, as all of the items seem to reflect substantial components of the shared general factor. However, the fit of this model was far from ideal. The resulting chi-square, with 27 degrees of freedom, equaled 156.49, a highly significant value. Additionally, the value of the Comparative Fit Index (Bentler, 1990) was only .78, indicating a relatively poor fit of model to the sample data.

[4] Actually the program COSAN (McDonald, 1978) was originally used to do the estimation. For the present chapter, however, the identical estimates were recomputed using PROC CALIS.



Residual Variances

ATD1	.37	ATD4	.39	ATD7	.50
ATD2	.15	ATD5	.56	ATD8	.28
ATD3	.42	ATD6	.17	ATD9	.31

Figure 14

Four-Factor ATD Model (From Figure 3 in Judd, Jessor, & Donovan, 1986)

A close inspection of the nine items plus an examination of the discrepancies between the correlation matrix predicted by the single factor model and the sample matrix suggested that the nine items might be better thought of as measuring four different, although related, constructs. The first two items seem to concern petty thievery; the next three dishonesty; items 6 and 7 fighting; and the last two items vandalism. Presumably there might exist related but slightly different attitudes in the four domains. Accordingly, a four factor solution was estimated, with each measured variable loading on one of four latent constructs and these latent constructs allowed to correlated with each other. The maximum likelihood estimates of the loadings and construct intercorrelations are given in Figure 14, which is reproduced from Figure 3 of Judd, Jessor, and Donovan (1986). The fit of this four factor model was considerably better than that of the one factor model, $\chi^2(21) = 26.08$; $\underline{p} > .20$; CFI = .99.

The single factor model is in fact nested under the four factor model, since the four factor model reduces to a single factor one if all six of the correlations among the factors are constrained to equal 1.0. Hence, we can ask not only about the fit of the four factor model in an absolute sense, but also about its comparative fit, relative to the single factor model. The chi-square difference between the two models equals 130.41 which is highly significant with 6 degrees of freedom.

In sum, this approach to the construct validity of these nine variables revealed that they are not consistent with a model in which the only thing they are assumed to have in common is a single underlying attitude towards deviant behavior. Rather, we must speak of a series of highly interrelated, although discriminable, attitudes, with the multiple measures of each one showing considerable convergent validity.

Example 2: Multitrait - Multimethod Matrix Specification. As the second example of the utility of confirmatory factor analysis in assessing construct validity, let us return to the data that we used to illustrate Generalizability Theory. Recall that in these simulated data 100 subjects rated their agreement with three items assessing prejudice levels towards three outgroups. The sums of squares and mean squares for these data were given in Table 12. The resulting multitrait - multimethod correlation matrix was presented in

Table 13. The nine variables that are correlated are subjects' responses to the three questions (methods) designed to tap prejudice towards each of the three outgroups (traits).

The most complete construct theory for a multitrait - multimethod matrix is one in which both traits and methods are specified as latent constructs and each measured variable is allowed to load on its respective trait and method factors. These trait and methods factors are assumed to correlate with each other to some extent. Finally, each measured variable is assumed to have some unique residual variance, uncorrelated with all other latent constructs.

A bit more formally, the construct theory or equation for each measured variable is as follows:

$$X_{jki} = \lambda_j \eta_{ji} + \lambda_k \eta_{ki} + \lambda_\varepsilon \varepsilon_{jki} \qquad (8)$$

In this equation, X_{jki} is the i^{th} subject's score on the measured variable expressing prejudice toward the j^{th} outgroup using the k^{th} question. Thus subscript j refers to trait and subscript k refers to method. η_{ji} and η_{ki} are latent trait and method factors that are allowed to be corrrelated with each other. ε_{jki} is residual or error variance unique to the specific variable.

Although this model would seem to be the most appropriate for the construct validity of measures in a multitrait - multimethod matrix, considerable difficulty has been encountered in estimating its parameters across a wide variety of different matrices (Kenny & Kashy, 1992; Marsh, 1989). Kenny and Kashy (1989) attribute these estimation difficulties to potential identification problems in the model, given that a more restricted version of the full model, one with equal loading coefficients, can be shown to be underidentified. Kenny and Kashy (1992) recommend that a considerably more restrictive model be estimated, one in which each measure is assumed to load on a latent trait construct and these are allowed to correlate. Additionally, the residual or error components of each measured variable are allowed to correlate with the residuals of other variables that share the same measurement method. They refer to this model as the "correlated uniqueness model."

A model that is actually identical to the "correlated uniqueness" model but that still involves both trait and method factors is a variant on the full model with the restrictions that the method factors (the η_{ki}'s in equation 8) correlate neither with each other nor with the trait factors. Thus each measured variable loads on both trait and method factors, but each of the three method factors is assumed to be uncorrelated with all other latent variables. It turns out that for many multitrait - multimethod matrices, including the present one, this restricted version of the full model is fully compatible with the data.

Using PROC CALIS in SAS, we estimated the parameters of this restricted model. The estimates from the resulting standardized solution, using the maximum likelihood criterion, are given in Table 15. The model was found to be consistent with the data, $\chi^2(15) = 21.98$; $\underline{p} > .10$; CFI = .99, and the estimates reveal considerable evidence of convergent and discriminant validity of these measures. First, all nine variables have large and highly reliable loadings on their respective trait constructs. Thus, all nine variables show considerable convergent validity as measures of prejudice toward the specific outgroup referenced. Second, there is discriminant validity among the three trait constructs, since the inter-trait correlations, although substantial, are considerably less than 1.0. It thus appears that prejudice against one outgroup is moderately, although far from perfectly, correlated with prejudice against another outgroup. Finally, the measures show only mild influence deriving from the measurement method used. Only the loadings on the first method factor differ reliably from zero. Thus, there is evidence of discriminant validity against measurement method.

Table 15

Restricted Multitrait - Multimethod Matrix Model for Data in Table 13

Loadings

	Trait Factors			Method Factors			
Variable	Trait 1	Trait 2	Trait 3	Method 1	Method 2	Method 3	Residual
G1M1	.74	0	0	.28	0	0	.61
G1M2	.88	0	0	0	.21	0	.59
G1M3	.80	0	0	0	0	.08	.59
G2M1	0	.86	0	.25	0	0	.45
G2M2	0	.88	0	0	.16	0	.44
G2M3	0	.79	0	0	0	.30	.53
G3M1	0	0	.79	.40	0	0	.47
G3M2	0	0	.88	0	-.25	0	.42
G3M3	0	0	.87	0	0	.42	.28

Factor Correlations

	Trait 1	Trait 2	Trait 3	Method 1	Method 2	Method 3
Trait 1	1.00					
Trait 2	.37	1.00				
Trait 3	.40	.33	1.00			
Method 1	0	0	0	1.00		
Method 2	0	0	0	0	1.00	
Method 3	0	0	0	0	0	1.00

Earlier we examined the construct validity of these simulated data by estimating the variance components due to subjects, measures, groups, and their interactions. We promised then that we would show the equivalence of that approach with the more general confirmatory factor analysis approach just presented. Our presentation here follows the insights of Kenny (1994), although we present a slightly different model specification than the one he presents for the confirmatory factor analysis model that is equivalent to the variance components approach.

The model whose parameter estimates are provided in Table 15 is already considerably constrained compared to the full multitrait - multimethod matrix model that seems most theoretically appropriate but has proven not to be estimable. Let us attach some additional constaints, however. Specifically, we will additionally assume the following:

1) All loadings of observed variables on the three trait factors, λ_j , are equal to each other;

2) All three correlations among the three latent trait factors are equal to each other;

3) All loadings of observed variables on the three method factors, λ_k, are equal to each other; and

4) All of the observed variables' residual variances are equal to each other.

Obviously these are a very strong set of assumptions about the model. In fact, they are sufficiently strong that they are equivalent to assuming, among other things, that all of the observed variables have equal variances. In order to evaluate the merit of this assumption, one needs to estimate the model's parameters using the variance / covariance matrix among the observed variables rather than their standardized correlation matrix. The reason for this is that through standardization one already has forced equal variances. But with this new constrained model, the imposition for the equal variance assumption comes from the model itself and so we wish to estimate that model with the variables in their raw metric.

The estimation of an unstandardized model involves estimating the variances and covariances of the latent factors rather than their loadings. All loading coefficients are constrained to equal 1.0, the variances of the three trait factors are constrained to equal each other; the covariances among the three trait factors are constrained to equal each other; the variances of the three method factors are constrained to equal each other; and finally the residual variances are all constrained to equal each other. In this highly constrained model, there are only four parameters that are estimated (Maximum likelihood estimates from PROC CALIS of SAS are included in parentheses:

1) the variance of the latent trait factors (22.29);

2) the variance of the latent method factors (1.49);

3) the covariance of the latent trait factors (7.80); and

4) the residual variance to the observed variables (8.35).

These estimated values are equivalent to some of the estimated variance components computed from these same data in Table 12. Thus, the variance of the latent trait factors is equal to the estimated variance component due to subjects plus the estimated variance component due to the subject by group interaction ($S^2_S + S^2_{SG}$). The variance of the latent trait factors is exactly equal to the estimated variance component due to the subjects by measures interaction (S^2_{SM}). And the residual variance is equal to the residual variance component (S^2_R). Additionally, if we convert the covariance between the latent trait factors to a correlation, by dividing their covariance by the variance of the trait factors:

$$\frac{7.80}{22.29} = .35$$

we get the same value for the discriminant validity coefficient among the three latent trait factors that we computed in the variance component approach.

In sum, although the variance components approach to the analysis of this multitrait - multimethod matrix is certainly useful, it is equivalent to an extremely constrained confirmatory factor analysis model in which it is assumed that all observed variables have equal variance, the variances and covariances of all underlying trait constructs are equal, and method influences on measures are all equivalent. It seems reasonable to estimate the convergent and discriminant validity of observed variables in a multitrait - multimethod matrix initially in a form that doesn't make these strong assumtpions. In this sense the confirmatory factor analysis approach is considerably more general.

Linkages Between Representational and Psychometric Approaches

As we noted in the introduction, it is remarkable how infrequently and inadequately the two measurement traditions discussed in the previous two sections have communicated with each other. Indeed, it is a rare treatise on measurement that even discusses both of them within the same book or chapter. Exceptions other than this chapter include Dawes and Smith (1985) and Himmelfarb (1993). Both Cliff (1989) and Fischer (1995), among others, have directly examined implications of the axiomatic approach for the psychometric approach. In this concluding section we identify some major differences between the two approaches and then explore two areas in which the gap might be bridged or in which the approaches might inform each other.

Some Differences Between the Approaches

Focus of Relationships. The focus of the axiomatic approach is the relationship between individual observations. For example, in the illustrations of Thurstonian and Fechnerian scaling above, the key was the empirical relationship of judging one item more or less liberal than another item. Axiomatic models use observed empirical relationships to make predictions about specific observations (as in the case of transitivity, for example). This leads to an almost exclusive focus on internal consistency within a set of such empirical relationships. In contrast, the focus of the psychometric approach is the relationship between variables. For example, in one of the illustrations of the psychometric approach above, the key was the covariance between items assessing prejudice in different ways. No individual observations are important; rather, the focus is the pattern of relationships among variables. Psychometric models make predictions about specific covariances or relationships between variables. This leads to an almost exclusive focus on what might be called external consistency.

A by-product of the different foci is that reliability is almost never assessed for representational measures whereas it is of paramount importance for psychometric measures. Below we explore how this difference in focus might be partially bridged by considering how axiomatic approaches might profitably examine covariances among representational measures. Doing so would allow assessment of reliability as well as external consistency.

Falsification versus Verification. Axiomatic models make powerful predictions about specific observations. As a consequence, such models are easily falsified. A single observation, particularly without an adequate error theory for axiomatic models, can falsify a model and foil the construction of a scale. For example, a clear failure of transitivity definitely rejects even an ordinal scale. No monotonic transformation (i.e., a differential stretching or shrinking of the scale), no change in assumptions about a probability distribution, and in general no changes about any assumptions can rescue the model. Such model rejections often identify powerful context effects that reveal the psychological processes underlying responses to social psychological stimuli. The disadvantage is that the ease of falsification makes axiomatic models so fragile that it is often too difficult to construct scales to be used to make predictions about other psychological constructs.

Psychometric models, on the other hand, are more difficult to falsify and relatively easy to verify. No single observation can possibly falsify a psychometric model, but there can be problematic items whose covariances with other items are not consistent. When such problematic items are encountered, it is common practice to discard them until a reasonable model results. Thus, it is usually possible to find a set of items for which the psychometric model is verified. Even if a confirmatory factor analysis revealed a bad fit, one would be left wondering if the rejection were due to an incorrect model or if it were perhaps due to non-linearities in the response scale, incorrect assumptions about a statistical distribution, etc. In other words, with the psychometric model, any falsifications are always weak.

Difficulty of Respondent's Task. The construction of representational measures poses relatively simple tasks to respondents. Individuals simply indicate which alternative they prefer or which item they think has more of some specified property. The judgment tasks are almost always relative comparisons. In contrast, psychometric approaches require, sometimes implicitly, respondents to assign numbers—to do the measuring themselves—to represent their responses to items. Social psychology is one of the few scientific disciplines that expects the entities it studies, rather than the scientists doing the studying, to assign the numbers. Furthermore, the psychometric approach involves absolute judgments of individual items, rather than relative comparisons. In general, absolute judgments are more difficult psychologically than relative judgments.

The psychometric approach assumes that respondents are sophisticated users of rating scales and that they can maintain consistent use of the response scale across many items. As a consequence, the psychometric approach can be used with a more restricted set of respondents. In particular, it is difficult to use the psychometric approach with the very young or the very old, with people in cultures where exposure to rating scales is rare, and with animals. On the other hand, the relatively simpler comparison tasks of choice or "which has more" of the axiomatic approach can be adapted to a much greater variety of respondents. For example, it is relativley easy to construct (or test for the existence of) an ordinal scale for a dog's preferences for treats by seeing which treats are eaten first from pairs of alternatives. With repeated trials, it is even easy to construct a Thurstonian or Fechnerian scale if there is any inconsistency in preference.

There are, however, other issues which make the axiomatic approach more difficult for respondents than the psychometric aproach. While the simple, relative judgment tasks of representative measurement are in general easy, the sheer number of judgments required for all the internal consistency checks can sometimes be numbing to respondents. This is especially true if replications are used to assess reliability. Also, with more sophisticated axiomatic models, the judgment tasks, even though they are relative, can be rather complex psychologicallly. For example, to test axioms for multidimensional scaling requires respondents to judge whether one pair of entities is more or less similar than another pair of entities. To do this task, the respondent must first determine which attributes are approproiate for judging similarity in this case, assess the locations of all four stimuli on those attributes, develop some rule for psychologically measuring the distance between the alternatives in terms of those attributes, and then report the greater (or least) distance for the two pairs of entities. For interesting social stimuli, this can be very difficult to do!

There is another subtle way in which the tasks used to construct representational measures can be difficult. Thurstonian and Fechnerian methods were derived from psychophysical methods that depend on confusions. Hence, attitude items used in these scaling methods must be close enough to be "confused" so that at least some people think, for example, that a is more liberal than b, while others think the reverse. If the items are too extreme, then such confusions or reversals are unlikely. Similarly, items are most useful in the unfolding model if their scale values fall between the ideal points of respondents. If there is a normal distribution of ideal points, then it is best to have a number of items fairly close together to each other in the middle of the scale in order to make discriminations among respondents. Although relative judgments are easier, making many relative judgments about items that are close together psychologically can be quite difficult. The psychometric approach, on the other hand, tends to favor extreme items that are very far apart. Deciding whether one agrees or disagrees with extreme items can often be easy, even though it is an absolute judgment. As another bridge between the two approaches we explore below in more detail the implications of the unfolding model for item covariances in the psychometric approach.

Implications of an Unfolding Model for Item Covariances

One of the strengths of the representational approach to measurement is that it makes explicit a theoretical model for data (Coombs, 1964). In this section, we wish to start with a given theoretical model, namely the unfolding model for preference data, and explore its implications for assessments of construct validity using the resulting pattern of item intercorrelations and confirmatory factor analysis. The goal is to examine the implications of the unfolding model of preference data for the psychometric approach to assessing the construct validity of the items towards which preferences have been expressed. Our approach bears resemblances to that of Coombs and Kao (1960; see also Coombs, 1964, Chapter 8).

Imagine that we have a set of ten attitude statements or items on each of two issues, say abortion rights and affirmative action. Imagine further that the ten items on each issue can be ordered from the most liberal viewpoint to the most conservative in the sentiments they express on the issue. We assume that the eight scale values of these statements on each issue are -5, -4, -3, -2, -1, 1, 2, 3, 4, and 5. (The metric for these scale values will become clear shortly when we describe the distribution of individual ideal points in the two dimensional preference space.) If these two dimensions are orthogonal to one another, then in a two-dimensional space they are at right angles and all items on one dimension have scale values of zero on the second. On the other hand, it is quite reasonable that these two dimensions are correlated, with statements or items that are liberal on one issue, implying liberal points of view on the other as well. For instance, if the correlation between the two dimensions was .40, then the appropriate two-dimensional representation of the items puts the two dimensions at an angle of approximately 66 degrees.

Now imagine that the we have a population of individual ideal points distributed in a bivariate normal manner ($\mu = 0$, $\sigma = 1$) throughout the two dimensional space defined by the two sets of items and their intercorrelation.[5] Each individual ideal point has two coordinates that give the individual's ideal preferences or positions on the two dimensions. And an individual's expressed preference or (dis)agreement for each of the items is

[5] The distribution of ideal points in the space defines the metric for the item scale values. For instance, an item with a scale value of 2 on one of the two dimensions is two standard deviations from the mean of all item and ideal point scale values on that dimension.

assumed to equal the Euclidean distance between the individual's ideal point in the two dimensional space and the location of each item (ten items for each of two dimensions) in that space. In theory, one then has a set of expressed preferences or agreements of each individual with each of the 20 items in the space and an expected matrix of item intercorrelations can be constructed from these.

Based on this multidimensional unfolding model of item preferences, we can then select subsets of these 20 items and examine the resulting correlation matrix among these items to assess their construct validity using confirmatory factor analytic procedures. According to the unfolding model, we know that the individual items can in fact be ordered along the two dimensions, that the preferences of the subjects on those dimensions is a perfect function of their Euclidean distances from those items, and that the correlation between the two dimensions is known. The estimated confirmatory factor analysis model treats the items as indicators of two latent factors and both the indicator loadings and the factor correlation are estimated.

We conducted a series of simulations to permit us to examine the ability of confirmatory factor analysis to recover the two dimensional preference unfolding model. In each trial of the simulations, we randomly sampled 150 individual ideal points from the full population of ideal points, computed preferences (Euclidean distances) of these individuals for all of the 20 items, computed intercorrelations among those preferences, and then estimated the parameters of the two-factor confirmatory factor analysis model for subsets of the items. Two factors varied between the simulation trials: the true correlations between the two dimensions in the space (.2, .4, and .6) and the scale extremity of the subset of items included in the confirmatory factor analysis. On one set of simulations, the four items with scale values of +5, and -5 were included. Other simulations used sets of four items having scale values of +4 and -4, +3 and -3, +2 and -2, and +1 and -1. At each level of these two crossed factors (true dimensional intercorrelation and extremity of items) 200 simulations were run, sampling anew 150 individual ideal points for each.[6]

Table 16 contains the results of these simulations. Each cell of the Table presents the results from each set of 200 simulation trials, with the two varied factors fixed at their row and column values. The four numbers that are included in each cell are medians across the 200 simulation trials of the goodness of fit χ^2 value, the CFI (Bentler, 1990), the estimated correlation between the factors, and the percentage of simulation trials where convergence was achieved and standard errors of estimates computed. Medians are presented rather than means due to the positive skew in the distributions of the estimates.

[6] The simulations were conducted using PROC CALIS in SAS. The code is available from the authors.

Table 16

Simulation Results:

Confirmatory Factor Analysis Given Indicators from a Two-Dimensional Unfolding Model

		Parameter Value of Correlation between the Two Factors			
Scale Value of Item Indicators		**.0**	**.2**	**.4**	**.6**
-5, +5	χ^2	1.06	1.21	6.18	24.30
	CFI	1.00	1.00	.99	.97
	r	-.03	.19	.38	.58
	Converged	49.0%	85.5%	96.5%	88.5%
-4, +4	χ^2	1.17	1.67	8.64	28.57
	CFI	1.00	.99	.99	.96
	r	-.05	.17	.37	.57
	Converged	38.5%	77.5%	93.0%	85.5%
-3, +3	χ^2	1.73	3.73	15.41	39.46
	CFI	.99	.99	.97	.93
	r	-.01	.06	.27	.52
	Converged	19.5%	56.5%	76.5%	79.0%
-2, +2	χ^2	12.25	9.71	30.28	65.63
	CFI	.96	.97	.92	.85
	r	-.01	.02	.03	.04
	Converged	0%	2.5%	7.0%	18.0%
-1, +1	χ^2	76.32	49.41	92.17	140.42
	CFI	.74	.78	.75	.68
	r	-.03	5.50	4.42	3.38
	Converged	0%	0%	0.5%	0.5%

If we assume that an unfolding model underlies preference judgments, then the results of these simulations are informative about the important role of item selection in the use of confirmatory factor analysis to uncover the latent structure of the items. As the extremity of the scale values of the items decreases, there is a fairly dramatic decrease in the quality of the model's fit and its parameter estimates. The quality of fit of the models, examining both the χ^2 and the CFI, becomes quite unacceptable with items that have scale values less than 2 standard deviations away from the mean. Additionally, all items yield attenuated estimates of the factor intercorrelation, and this attenuation increases with less extreme items. Finally, there are considerable convergence problems with models with relatively non-extreme items. This is because such models, especially when the two dimensions are uncorrelated, tend to be empirically unidentified (Kenny, Kashy, & Bolger, in press). The lesson, then, is that if the Euclidean multidimensional unfolding model of preferences is an appropriate model, then confirmatory factor analyses of item intercorrelations yield satisfactory results only in the case of items that have scale values considerably more extreme than the ideal points of typical respondents.

Conceptually, these results derive from the fact that the magnitude of disagreement with more moderate items fails to indicate the direction of an individual's ideal point. For example, imagine two individuals with unidimensional ideal point values of +2 and -2. Although they dramatically disagree with each other, they are in agreement about their preferences for an item having a scale value of 0. Only items that are more extreme than respondents permit an unambiguous assessment of directional preferences.

Of course this extremity criterion for item selection poses the additional dilemma of insuring adequate variance in the actual measurement of subjects' preferences. In our simulations, the preferences (Euclidean distances) are continuously measured. Hence, regardless of the extremity of an item, there remains variation in preferences. If quite extreme items are used in actual preference data collection, where subjects indicate preferences on rating scales, one needs to be confident that variation in preferences will be picked up even with quite extreme items where the dominant response may be "disagreement."

Examining Covariances among Representational Measures

The strengths of the representational approach to measurement derive from the explicit theoretical models that underlie measurement and the internal consistency checks that can be used to evaluate the appropriateness of those theoretical models. But a consequence of these internal consistency checks is that external criteria for evaluating the utility of a scale or variable have generally not been used. In other words, those who have advocated representational approaches to measurement in social psychology have seldom encouraged researchers to exploit the virtues of the psychometric approach in evaluating the predictive success of those measures. Useful and valid measures ideally should meet not only the representational criteria (i.e., have an explicit theory that is internally testable), but they also ought to correlate with other measures of similar and related constructs in predictable ways, following the insights of the psychometric approach. To our mind (and in agreement with Dawes, 1994), a significant shortcoming of the representational approach is that issues of predictability, reliability, and external validation have generally been ignored.

Suppose, for instance, that we asked subjects to rank order attitude statements on an issue, say abortion rights, and from these rank orders, we applied unidimensional unfolding procedures, revealing scale values for both the atttitude statements and the subjects. And suppose further that the internal consistency checks of the unfolding approach were found to hold, so that we had confidence that a single attitude continuum was adequate to represent preferences in this domain. Would that be sufficient? What about the psychometric crieria for evaluating the reliability and construct validity of this measurement approach? Are these criteria irrelevant given that the internal consistency checks have been met?

We would like to suggest that these psychometric criteria are perfectly appropriate and that one of the shortcomings of the representational approach to measurement is that it has ignored the potential informativeness of covariances at the aggregate level. It would certainly be informative to know, for instance, whether these subjects' scale values show evidence of reliability, stability over time, and convergent and discriminant validity with other measures of the same and different constructs. One could imagine, for instance, embedding these scale scores as one variable in a multitrait-multimethod study of different methods of measuring two or more attitudes and then using confirmatory factor analysis to evaluate the discriminant and convergent validity of the included measures. We suspect that the assurance from the unfolding model that there is in fact a single underlying dimension that is assessed ought to mean that these scale scores should show evidence of considerable construct validity when patterns of covariance are examined. But in fact, the empirical work to demonstrate this convergence has not been done.

References

Aguinis, H., Pierce, C.A., & Quigley, B.H. (1995). Enhancing the validity of self-reported alcohol and marijuana consumption using a bogus pipeline procedure: A meta-analytic review. Basic and Applied Social Psychology, 16, 515-534.

Anderson, N.H. (1981). Foundations of information integration theory. New York: Academic Press.

Anderson, N.H. (1982). Methods of information integration theory. New York: Academic Press.

Anderson, N.H., Ed. (1991). Contributions to information integration theory, Vols 1-3. Hillsdale, NJ: Lawrence Erlbaum Associates.

American Psychological Association. (166). Standards for educational and psychologcial tests and manuals. Washington, D.C.: American Psychological Association.

Anastasi, A. (1961). Psychological testing (2nd ed.). New York: Macmillan.

Armour, D.J. (1974). Theta reliability and factor scaling. In H. L. Costner (Ed.), Sociological methodology, 1973-1974 (pp. 17-50). San Francisco: Jossey-Bass.

Arrow, K.J. (1951). Social choice and individual values. New York: Wiley.

Bennett, J.F., & Hays, W.L. (1960). Multidimensional unfolding: Determining the dimensionality of ranked preference data. Psychometrika, 25, 27-43.

Bentler, P.M. (1968). Alpha-maximized factor analysis and its relation to alpha and connonical factor analysis. Psychometrika, 33, 335-346.

Bentler, P.M. (1979). Mulivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 31, 419-456.

Bentler, P.M. (1990). Comparative fit indices in structural models. Psychological Bulletin, 107, 238-246.

Bentler, P.M. (1993). EQS prgroam manual. Los Angeles: BMD Statistical Software.

Bernieri, F.J., & Rosenthal, R. (1991) Interpesonal corrdination: Bheavior matching and interactional synchrony. In R. S. Felman & B. Rimé (Eds.), Fundamentals of nonverbal behavior (pp. 401-432). Cambridge: Cambridge University Press.

Bollen, K., & Long, S. (1993). Testing structural equation models. Newbury Park, DA: Sage.

Bradley, R.A., & Terry, M.E. (1952). Rank analysis of incomplete blck designs: I. The method of paried comparisons. Biometrika, 39, 324-245.

Burke, C.J., & Zinnes, J.L. A paired comparison of pair comparisons. Journal of Mathematical Psychology, 2, 53-76.

Cacioppo, J.T., & Petty, R.E. (1979). Attitudes and cognitive response: An electrophysiological approach. Journal of Personality and Social Psychology, 37, 2181-2199.

Campbell, D.T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. American Psychologist, 15, 546-553.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, 68, 361-368.

Chernoff, H., & Rizvi, H.M. (1975). Effect on classification error of random permutations of features in representing multivariate data by faces. Journal of the American Statistical Association, 70, 548-554.

Clark, W.A.V. (1982). A revealed prerence analysis of intraurban migration choices. In R.G. Golledge & J.N. Rayner (Eds.), Proximity and preference: Problems in the multidimensional analysis of large data sets. Minneapolis: University of Minnesota Press.

Cleveland, W.S. (1933a). A model for studying display methods of statistical graphics (with discussion). Journal of Computational and Statistical Graphics, 2, 323-343.

Cleveland, W.S. (1983b). Visualizing data. Summit, NJ: Hobart Press.

Cleveland, W.S., & McGill, R. Graphical perception: The visual decoding of quantitative information on grpahical displays of data. Journal of the Royal Statistical Soceity, Series A, 150, 192-229.

Cliff, N. (1989). Ordinal consistency and ordinal true scores. Psychometrika, 54, 75-91.

Cliff, N.F. (1992). Abstract measurement theory and the revolution that never happened. Psychological Science, 3, 186-190.

Condorcet, M. (1785). Essai sur l'application de l'analyse a la probabilité des decisions rendues a la pluralité des voix. Paris: Impr. Royale. [reprinted New York: Chelsea Publishing Co., 1972]

Coombs, C.H. (1950). Psychological scaling without a unit of measurement. Psychological Review, 57, 145-158.

Coombs, C.H. (1964). A theory of data. New York: John Wiley and Sons.

Coombs, C.H., & Avrunin , G.S. (1988). The structure of conflict. Hillsdale, NJ: Lawrence Erlbaum Associates.

Coombs, C.H., Coombs, L.C., & McClelland, G.H. (1975). Preference scales for number and sex of children. Population Studies, 29, 273-298.

Coombs, C.H., & Kao, R.C. (1960). On a connection between factor analysis and multidimensional unfolding. Psychometrika, 25, 219-231.

Cook, S.W., & Selltiz, C. (1964). A multiple-indicator approach to attitude measruement. Psychological Bulletin, 62, 36-55.

Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Cornfield, J., & Tukey, J.W. (1956). Average values of mean squares in factorials. Annals of Mathematical Statistics, 27, 907-949.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Cronbach, L.J. (1960). Essentials of psychological testing (2nd ed.). New York: Harper.

Cronbach, L.J. (1984). Essentials of psychological testing (4th ed.). New York: Harper.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Dawes, R.M. (1994). Psychological measurement. Psychological Review, 101, 278-281.

Dawes, R.M., & Moore, M. (1979). Guttman scaling orthodox and randomized responses. In F. Peterman (Ed.), Attitude Measurement.(pp. 117-133). Gottinger: Verlag für psychologie.

Dawes, R.M., & Smith, T.L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), The handbook of social psychology, Vol. 1 (pp. 509-566). New York: Random House.

Dawkins, R. (1969). A threshold model of choice behavior. Animal Behavior, 17, 120-133.

de Leeuw, J., Young, F.W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. Psychometrika, 41, 471-503.

Dovidio, J. F., Evans, N., & Tyler, R. (1986). Racial stereotypes: The contents of their cognitive representations. Journal of Experimental Social Psychology, 22, 22-37.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. Journal of Personality and Social Psychology, 50, 229-238.

Fischer, G.H. (1995). Some neglected problems in IRT. Psychometrika, 60, 459-487.

Fyans, L.J., Jr. (Ed.) (1983). Generalizability theory: Inferences and practical applications. San Franscisco: Jossey-Bass.

Green, P.E., & Srinivasan, V. (1990). Conjoint analysis in marketing: new developoment with implications for research and practice. Journal of Marketing, 54, 3-19.

Greenberg, B.C. Abdula, A.L. Simmons, W.R., & Horvitz, D.G. (1969). The unrelated question in randomized response model, theoretical framework. Journal of the American Statistical Association, 64, 520-539.

Hays, W.L., & Bennett, J.F. (1961). Multidimensional unfolding: Determining configuration from complete rank order preference data. Psychometria, 26, 221-238.

Hastie, T.J. (1992). Generalized additive models. In J.M. Chambers & T.J. Hastie (Eds.), Statistical models in S. Pacific Grova, CA: Wadsworth.

Hastie, T.J., & Tibshirani, R. (1990). Generalized additive models. London: Chapman & Hall.

Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.) The psychology of attitudes (pp. 23-88). Fort Worth, TX: Harcourt, Brace, Jovanovich.

Jones, E.E., & Sigall, H. (1977). The bogus pipeline: A new paradigm for emasuring affect and attitude. Psychological Bulletin, 76, 349-364.

Jöreskog, K.G. (1981) Analysis of covariance structures. Scandinavian Journal of Statistics, 8, 65-92.

Jöreskog, K.G., & Sörbom, D. (1993). LISREL8: The SIMPLIS command language. Chicago: Scientific Software.

Jaccard, J., & Wan, C.K. (1995). Measurment error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. Psychological Bulletin, 117, 348-357.

Jessor, R. Donovan, J.E., & Costa, F.M. (1991). Beyond adolescence: Problem behavior and young adult development. New York: Cambridge University Press.

Jessor, R., & Jessor, S.L. (1977). Problem behavior and psychosocial development: A longitudinal study of youth. New York: Academic Press.

Judd, C.M., Jessor, R., & Donovan, J.E. (1986). Structural equation models and personality research. Journal of Personality, 54, 149-198.

Judd, C.M., & Kenny, D.A. (1981). Estimating the effects of social interventions. New York: Cambridge University Press.

Kenny, D.A. (1979) Correlation and causality. New York: Wiley-Interscience.

Kenny, D.A. (1994). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In P.E. Shrout & S.T. Fiske (Eds.). Personality, research, methods, and thoery (pp. 111-124). Hillsdale, NJ: Erlbaum.

Kenny, D.A., & Judd, C.M. (in press). A general procedure for the estimation of interdependence. Psychological Bulletin.

Kenny, D.A., & Kashy, D.A. (1992). The analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin, 112, 165-172.

Kenny, D.A., Kashy, D.A., & Bolger, N. (in press). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.). The Handbook of Social Psychology.

Kirk, R.E. (1982). Experimental design. Belmont, CA: Brooks/Cole.

Kosslyn, S.M. (1989). Understanding charts and graphs. Applied Cognitive Psychology, 3, 185-226.

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). Foundations of measurement, Vol. 1. Additive and polynomial representations. New York: Academic Press.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. Journal of the Royal Statistical Society, Series B, 27, 251-263.

Kruskal, J.B., & Wish, M. (1978). Multidimensional scaling. Beverly Hills & London: Sage.

Lerner, P.E., & Noma, E. (1980). A new solution to the problem of finding all numerical solutions to ordered metric structures. Psychometrika, 45, 135-137.

Loehlin, J.C. (1992). Latent variable models: An intorduction to factor, path, and structural analysis. Hillsdale, NJ: Erlbaum.

Luce, R.D. (1959). Individual choice behavior. New York: Wiley.

Luce, R.D. (1995). Four tensions concerning mathematical modeling in psychology. Annual Review of Psychology, 46, 1-26.

Luce, R.D., Krantz, D.H., Suppes, P., & Tversky, A. (1990). Foundations of measurement. Vol. 3. Representation, axiomatization, and invariance. New York: Academic Press.

Macrae, C.N., Bodenhausen, G.V., Milne, A.B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. Journal of Personality and Social Psychology, 67, 808-817.

Marley, A.A.J. (1992). Measurment, models, and autonomous agents. Psychological Science, 3, 93-96.

Marsh, H.W. (1989). Confirmatory factor analyses of mutlitrait-multimethod data: Many problems and a few solutions. Applied Psychological Measurement, 13, 335-361.

McClelland, G.H., & Coombs, C.H. (1975). ORDMET: A general algorithm for constructing all numerical solutions to ordered metric structure. Psychometrika, 40, 269-290.

McDonald, R.P. (1978) A simple comprehensive model for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 31, 59-72.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational Measurement (pp. 13- 102). New York: Macmillan Publishing Co.

Messick, S. (1995). Validation of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Michell, J. (1990). An introduciton to the logic of psychological measurement. Hillsdal, NJ: Lawrence Erlbaum Associates.

Nickerson, C. A., McClelland, G. H., & Petersen, D. M. (1991). Measuring contraceptive values: An alternative approach. Journal of Behavioral Medicine, 14, 241-266.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). The measurement of meaning. Urbana: University of Illinois Press.

Petty, R.E., & Cacioppo, J.T. (1983). The role of bodily responses in attitude measurment and change. In J.T. Cacioppo & R.E. Petty (Eds.), Social psychphysiology: A sourcebook (pp. 51-101). New York: Guilford Press.

Pfanzagl, J. in collaboration with V. Baumann & H. Huber. (1968). Theory of measurement. New York: Wiley.

Rankin, R.E., & Campbell, D.T. (1955). Galvanic skin response to Negro and white experimenters. Journal of Abnormal and Social Psychology, 51, 30-33.

Roskam, E.E. (1992). ORDMET3: An improved algorithm to find the maximin solution to a system of linear (in)equalities. Methodika, 6, 30-53.

SAS Institute. (1990). SAS/STAT user's guide, Version 6, Fourth edition (Vols. 1 & 2). Cary, NC: SAS Institute.

Shah, P., & Carpenter, P.A. (1995). Conceptual limitations in comprehending line graphs. Journal of Experimental Psychology: General, 124, 43-61.

Shrout, P.E., & Fleiss, J.L. (1979) Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. Journal of the American Statistical Association, 82, 454-465.

Stevens, SS. Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), Handbook of experimental psychology (pp. 1-49). New York: Wiley.

Suppes, P. Krantz, D.H., Luce, R.D., & Tversky, A. (1989). Foundation of measurement. Vol. 2. Geometrical, threshold, and probabiliistic representations. New York: Academic Press.

Suppes, P., & Zinnes, J.L. (1963). Basic measurement theory. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), Handbook of mathematical psychology (Vol. 1, pp. 1-76). New York: Wiley.

Thurstone, L.L. (1927). A law of comparative judgment. Psychological Review, 34, 373-286.

Torgerson, W.S. (1958). Theory and methods of scaling. New York: Wiley.

Tufte, E.R. (1990). Envisioning information. Cheshire, CT: Grpahics Press.

Tversky, A. (1969). Intransitivity of preferences. Psychological Review, 76, 31-48.

Tversky, A., & Russo, J.E. (1969). Substitutaqbility and similarity in binary choices. Journal of Mathematical Psychology, 6, 1-12.

Tversky, A., Slovic, P., & Sattath, S. (1988). Contingent weighting in judgment and choice. Psychological Review, 95, 371-384.

Tversky, B., & Schiano, D.J. (1989). Perceptual and conceptual factors in distortions in memory for graphs and maps. Journal of Experimental Psychology: General, 118, 387-398.

van der Ven, A.H.G.S. (1980). Introduction to scaling. New York: Wiley.

Wainer, H., & Theissen, D. (1981). Graphical data analysis. Annual Review of Psychology, 32, 191-241.

Warner, S.L. (1965) Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63-69.

Webb, E.J., Campgell, D.T., Schwartz, R.D., & Sechrest, L. (1966). Unobtrusive measures: Non-reactive research in the social sciences. Chicago: Rand McNally.

Winer, B.J., Brown, D.R., & Michels, K.M. (1991). Statistical principles in experimental design. New York: McGraw-Hill.

Wish, M. (1971). Individual differences in perceptions and preferences among antions." In C.W. King & D. Tigert (Eds.), Attitude reserach reaches new heights. Chicago: American Marketing Association.

Wittenbrink, B., Judd, C.M., & Park, B. (in press). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. Journal of Personality and Social Psychology.

Wittink, D.R., & Cattin, P. (1989). Commercial use of conjoint analysis: An update. Journal of Marketing, 53, 91-96.

Woodmansee, J.J. (1970). The pupil response as a measure of social attitudes. In G.F. Summers (Ed.), Attitude measurement (pp. 514-533). Chicago: Rand McNally.

Yellott, J.I. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. Journal of Mathematical Psychology, 15, 109-144.

Young, F.W. (1984). Scaling. Annual Review of Psychology, 35, 55-81.