1

Scaling Theory

A facet of measurement not covered in the main text is that of scaling theory. Scaling theory focuses on the mathematical functions by which the qualities or properties of a construct map onto the measurement scale designed to measure those qualities and properties. Steven's taxonomy of nominal, ordinal, interval, and ratio scales, discussed in Chapter 13 of the main text, is an example of a scaling theory, albeit a crude one. This primer introduces examples of scaling theories, but with an emphasis on creative theory construction at the level of measurement. It is for those who are more quantitatively inclined, but it is introductory and stresses the conceptual underpinnings of scaling theory. The goal is to illustrate to you some of the creative ways that measurement theorists have approached the process of measurement.

We begin with a discussion of multi-item scales and introduce the theoretical concept of tracelines, also called item operating characteristics. We show how different theories about tracelines lead to different scaling strategies. This leads to a brief characterization of the popular item response theory (IRT) approach to scale construction, which builds on the concept of item operating characteristics. To illustrate the creativity that psychometricians have brought to scaling theory, we next discuss two scaling approaches, the first grounded in conjoint and functional measurement and the second in multidimensional scaling. Conjoint measurement is widely used in marketing and has its roots in the influential measurement work of Luce and Tukey (1964). A variant of it used in psychology is functional measurement (Anderson, 1981, 1982). The approaches illustrate how measurement theorists have tackled difficult scaling problems in highly creative ways for the analysis of human judgment and decision making. We then describe multidimensional scaling, a scaling theory inspired by maps that use the Cartesian coordinate system. The idea behind this approach is that just as we have physical maps to show the location of geographic points of interest based on the dimensions of north-south and east-west, people often have mental maps of "objects" (e.g., people, events, places) that locate those objects in a psychological coordinate system. The task is to discover and represent those mental maps.

THE CONSTRUCTION OF MULTI-ITEM SCALES

In this section, we discuss issues for the construction of multi-item scales, focusing

primarily on matters of theory. We first focus on the general practice of item screening from a more mundane practical level and then address theory-based item screening using item operating characteristics. It is the latter screening approaches where you will see creative theorizing in action.

Item Generation and Screening: General Considerations

When constructing a scale, the first step is to clearly define the construct in question and to map out the conceptual domain that needs to be represented in order to be faithful to the core elements of the construct. For example, if social support is conceptualized as having four distinct facets (information support, tangible support, emotional support, and companionship support), then one will want to generate a set of items that taps into each facet. In this sense, theory typically guides item generation. Chapters 5 and Chapter 13 in the main text detail the importance and fundamentals of such concept mapping.

When generating items, one wants to maximize the reliability and validity of each item as an indicator of the construct to be measured. Reliability means item responses are free of random error. Validity implies item response are not also biased by systematic error. As noted in Chapter 13, many factors can affect item error variance and you need to take these factors into account as you generate your items. It is common practice to generate far more items than one intends to include on the final scale because, invariably, some items will be ill-behaved in terms of reliability and validity and will need to be rejected for scale inclusion when empirically evaluated.

A response metric also must be chosen for each item, such as a two point agree-disagree format or a five-point frequency format. Chapter 14 in the main text discussed factors to consider when choosing a response metric. In general, it is better to have item metrics that are more precise and where adverb qualifiers have been carefully chosen to approximate interval level properties, but this varies by the type of scaling theory used.

Typically, one will conduct a psychometric study after initial item generation to screen out poorly performing items or to flag items where wording revisions are necessary. A useful strategy is to conduct, if possible, a test-retest study in which the same individuals respond to items at two different points in time so that response consistency across the two assessments can be determined. Inconsistent responses between the assessments implies the item is susceptible to random error. The time interval between the assessments should not be too long because if it is, the construct may change over time. One then will not know if the inconsistent responses are due to random error or to the fact that the construct has changed. You typically will want to select a time interval in which you are confident the construct does not change. Too short a time interval also is a danger if respondents then try to recall what their responses were at the

prior assessment. As well, people might become irritated when asked to respond to the same items twice. In our instructional sets for the second assessment, we tell respondents that good scientific practice is to determine how people respond to the same items on different occasions and that they (the respondents) should respond to each item honestly and based on how they feel now, without trying to recall how they responded in the past. We find most people are understanding if we are transparent with them about our purposes. We typically use a one or two week test-retest interval. Items that show unacceptable levels of response consistency are then eliminated or revised.

In addition to response consistency, a second item property one can assess in the test-retest study is the response base rate for each item. Suppose we use a two point response metric for items, agree-disagree. If 90% or more of respondents agree with an item (or 90% or more disagree with it), then the item is of questionable utility for measurement purposes. The idea is that you are trying to measure variation in the underlying construct of interest based on the assumption that there is meaningful variation in it. If the response to an item shows little variability, then how can it be sensitive to the variability in the underlying construct? It can't. As such, we either eliminate items that have base rate problems (i.e., show highly skewed response patterns) or we change the wording of them so that responses become more variable, perhaps by making the wording for the item more or less extreme.

As an example, an item measuring attitudes towards getting pregnant in female middle school adolescents might read "My getting pregnant at this time in my life would be bad," to which respondents either agree or disagree with it. This will be a poor item psychometrically because almost all middle school females will agree with it. It does not tap into the extant variability in how "bad" youth perceive a pregnancy at this time in their lives to be. By making the statement more extreme, we might observe response patterns that better reflect such variability, such as "My getting pregnant at this time in my life would be one of the worst things that could possibly happen to me." Base rates also are relevant for items with more than two response categories. The concern is with an unsatisfactory bunching of scores at one end of the distribution.

In the test-retest study, it also may be useful to include measures of response sets and response bias, as discussed in Chapter 13. For example, measures have been developed to assess (1) social desirable response bias (the tendency to respond to items so as to create a positive impression rather than reflecting one's true opinion), (2) acquiescence response bias (the tendency to endorse items/questions, independent of their content), (3) disacquiescence response bias (the tendency to disagree with items independent of their content), (4) extreme response bias (the tendency to use the extremes of a rating scale independent of item/question content), (5) midpoint response bias (the

tendency to use the midpoint of a rating scale independent of item/question content) and (6) non-contingent response bias (the tendency to respond to items carelessly, randomly, or non-purposefully). Although these tendencies are thought to be general characteristics of individuals, it is possible that certain items are more likely to elicit such response sets than others. Items that show moderate to strong correlations with these artifacts might be screened out or revised; see Baumgartner and Steenkamp (2001) and Stoeber (2001).

The test-retest study also can be used to evaluate items for their concurrent or construct validity. This involves correlating item responses with measures of other constructs that the target construct is thought to be correlated with. For example, if you are developing a measure of school connectedness among youth, a large body of research has established that a moderate correlation between school connectedness and grade point average (GPA) exists. One could include a measure of GPA and then correlate each item with that index. Items with weak relationships to GPA might be screened out or revised.

It is usually good psychometric practice to evaluate the above properties for different subgroups within the test-retest study to ensure that the items perform well across subgroups, i.e., that the item properties generalize across different subpopulations. There are elegant methods in structural equation modeling that can be applied to explore metric generalizability across groups and time (see Kline, 2016).

When conducting the test-retest study, it is important to ensure you have sufficient sample size to counter the presence of sampling error in your psychometric evaluations. Traditionally, social scientists base sample size decisions on statistical power, but for psychometric studies it is best to also select sample sizes based on desired margins of error or confidence interval width. For example, we would expect reliability estimates for an item to be relatively large, ideally yielding test-retest correlations in the 0.70 range. Significance tests for such large correlations are not very meaningful. Rather, we want our estimates of item reliability in a population to be within a certain margin of error (MOE), such as plus or minus 0.05 correlation units. There are methods for determining sample sizes necessary to achieve desired MOEs. These are available in the ASA statistical package described on our website under the “tutorials/videos” tab.

Finally, once ill-behaved items have been screened or revised, it is useful to subject the remaining items to cognitive response testing per Chapter 13. At the conclusion of this first screening task, you will want to make sure you still have a sufficient number of items in each relevant domain of your construct so that the construct remains adequately mapped. This is important because the next screening step will eliminate yet more items.

Item Screening using Item Operating Characteristics

After you have screened out poor items based on the initial pilot work, an equally

important screening step involves using a formal scaling theory and empirically testing items to ensure they are consistent with that theory. Most scaling theories make use of a concept called an ***item operating characteristic*** (IOC), also called a ***traceline*** (Green, 1954). An IOC refers to the relationship between item endorsement and a person's true location on the underlying dimension of the measured construct, i.e., his or her true score. To be concrete, suppose we have 4 items that are thought to measure political attitudes reflecting social conservatism, namely the tendency to embrace social policies that reflect conservative as opposed to liberal thinking. Conceptually, the true underlying social conservatism construct is thought to impact how individuals respond to the items. Suppose each item has two response options, agree or disagree, and endorsement of a given item reflects a more conservative attitude towards social policies. We can, in theory, plot the relationship between the probability of endorsing a given item and peoples' true scores on the dimension. Figure 1.1 presents the IOC assumed by many popular scaling approaches, namely a linear IOC; the more socially conservative individuals are, the more likely they are to endorse socially conservative items and the less likely they are to endorse liberal items. This IOC is evaluated is by correlating each item with the total attitude score represented by the sum of all the items retained after the initial screening (with appropriate item reverse scoring). The sum of the items represents an (imperfect) proxy for the person's true score on the underlying dimension. Items are then eliminated that fail to conform to this IOC. An item-total correlation less than 0.40 is generally seen as suspect. These correlations can be evaluated using data from the test-retest study or in a separate psychometric study designed for IOC analytic purposes.

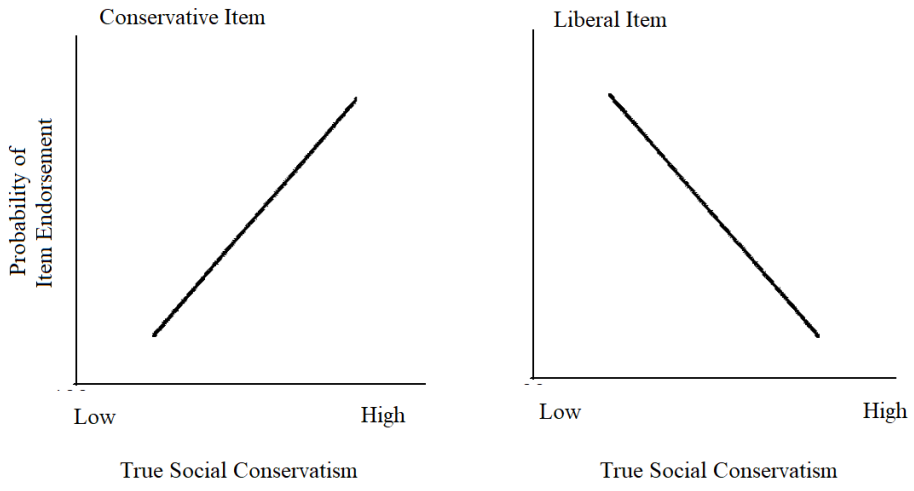

FIGURE 1.1. Linear Traceline

There are different methods for empirically evaluating IOCs depending on the IOC assumed. All of the methods are approximate because to unambiguously test an IOC, we need to know the true scores of individuals on the underlying dimension. Most methods either use proxies for the true scores, per the above example for linear tracelines, or they make assumptions about the true scores that permit formal tests. For a discussion of item selection methods using tracelines, see Green (1954), Edwards (1957), Lord (1980) and Meade and Meade (2010).

Psychometricians have specified other types of possible IOCs than linear ones. These other types might yield scales that are better suited to your research questions. One such approach is that of Guttman (1944) scaling. The logic of Guttman scaling is easiest to understand with reference to a test of math ability. Each item on the test might have a different level of difficulty in terms of the ability required to solve it (also called an item's *scale value*). Suppose, for the sake of illustration, we let the degree of difficulty of an item be characterized on a 0 to 10 scale, where 0 is very easy, 10 is very difficult, and the higher the number, the more difficult the item. For a Guttman scale, if a person's true math ability exceeds the difficulty level of the item, the probability the person will get the item correct is 1.0; if the item difficulty exceeds the person's math ability, then the probability the person will get the item correct is 0.0. This dynamic yields a step-shaped IOC rather than a linear one, as illustrated in Figure 1.2 for an easy, moderately difficult, and a difficult item.

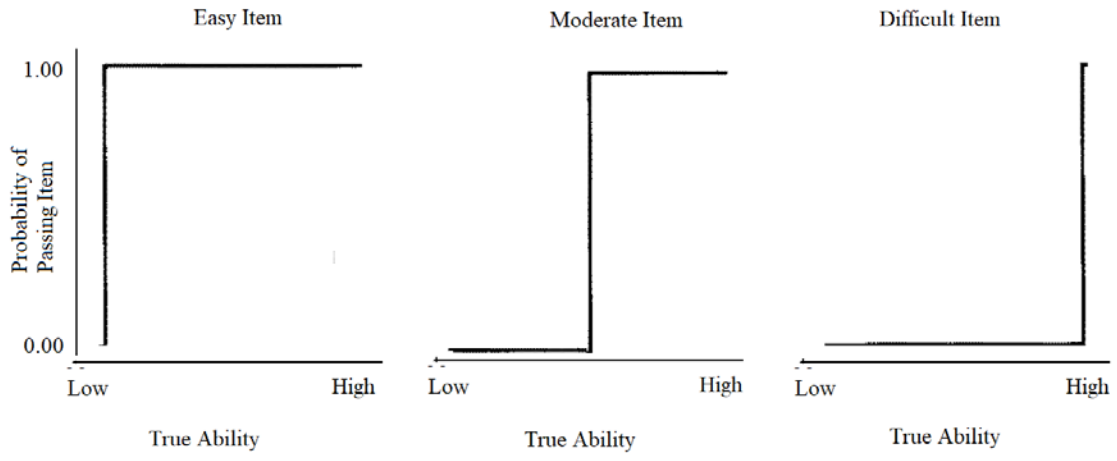

FIGURE 1.2. Step-Shaped Traceline

In Guttman scaling, items whose scale values or difficulty levels are thought to

span the range of the underlying dimension are identified and then the items are formally tested for a step-shaped IOC. Items that do not have this property are discarded. This approach can be used for any dimension of interest, not just ability dimensions. Examples include scaling the difficulty of everyday activities performed by the elderly (Rosow, & Breslau, 1966), scaling adolescent transitions to substance use (from alcohol use to cigarette use to marijuana use to hard drugs; see Andrews, Hops, Ary et al. 1991), stages of courtship or relationships (King & Christensen, 1983), and the measurement of condom use skills for HIV prevention (Lindemann, 2003). An interesting application in anthropology by Kay (1964) scaled the ownership of consumer goods for households in a small township in French Polynesia. Kay found the following goods conformed to a step shaped IOC for an asset-based index of SES (expressed here in ordered diagnosticity): a primus stove, a bicycle, a radio, a two-wheeled motor vehicle, a gas stove, a refrigerator, and an automobile. For example, a household that owned a radio also owned a bicycle and a primus stove but if it did not have a two-wheeled motor vehicle, it likely did not have a gas stove, a refrigerator or an automobile. Might a Guttman scale fit your research questions? Do you really want to just mindlessly default to a linear IOC?

Another scaling approach that does not use linear tracelines is based on the work of Thurstone (1928). In applying his method of equal appearing intervals, for example, Thurstone generated a pool of attitudinal items that he felt spanned the dimension of unfavorable to favorable statements about the target attitude object, such as attitudes about environmental conservation. His first task was to specify the location of each item on the underlying evaluative dimension (e.g., a moderately unfavorable item, a neutral item, a strongly favorable item), i.e., to identify its scale value. Thurstone initially used a panel of expert judges to determine each item's scale value, but later developed methods for identifying them based on psychophysical scaling methods (see Edwards, 1957; Edwards & Gonzalez, 1993). The details need not concern us here. Suffice it to say the methods were used to identify the scale value for each item. If the underlying dimension was on a metric from, say, -5 to +5, one item might have a scale value of -5.0, another item might have a scale value of -4.5 on the dimension, and so on, up through one or more items with highly positive scale values.

The goal of Thurstone's approach was to identify a set of approximately 10 to 15 items that spanned the dimension of interest and that had more or less equally spaced scale values (to yield an interval level scale). Items identified as having ambiguous scale values during the scale value estimation process were eliminated. For items that were retained, Thurstone applied a second screening criterion, namely whether the item had a theoretically appropriate IOC. Thurstone made the assumption that an item with a given scale value should be most likely to be endorsed by individuals whose attitudes were

located at the same position on the attitude dimension as the item. The greater the discrepancy between the person's true location on the dimension and the item's scale value, the lower the probability the person should be to endorse the item. For example if a person is slightly negative towards environmental conservation, then s/he should be most likely to endorse items that are slightly negative and reject items that are more extreme in either direction because the items are “too favorable” towards conservation or “too unfavorable” towards conservation. This IOC is shown for three different items in Figure 1.3. Note that for an item with a scale value that is relatively neutral, the IOC is non-monotonic. For items with more extreme scale values on either end of the dimension, the IOC is approximately linear.

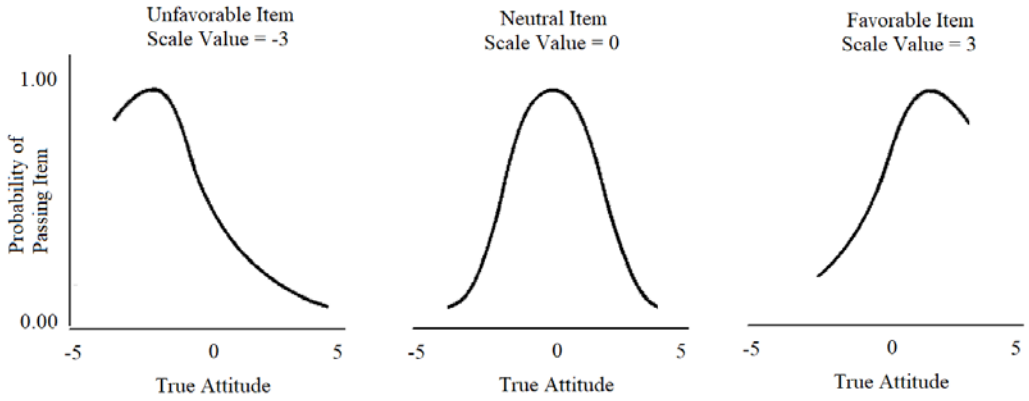

FIGURE 1.3. Ideal Point Traceline

Thurstone’s presumed IOC has been referred to in the psychometric literature as an *ideal point model* (Drasgow, Chernyshenko & Stark, 2010). If an item does not conform to this IOC, it is eliminated from the scale. For the final items, a person’s attitude score is the mean scale value of all items endorsed. If a person endorses three items with scale values of -2.6, -3.0, and -3.4, the overall attitude score is -3.0.

Thurstone’s methods were noteworthy because they were thought to yield approximately interval level metrics and were tied to widely accepted psychophysical principles of the time. His methods can be applied to dimensions other than attitudinal, such as personality scales or other diverse judgment dimensions. Drasgow et al. (2010) argue that Thurstone’s assumed IOC is more representative of how people make cognitive judgments about items/statements in attitude or opinion surveys and because of this, are preferred. People essentially ask themselves, the argument goes, ‘‘does this

statement closely describe my viewpoint?'' and, if so, they endorse it. Drasgow et al. also argued that Thurstone's IOC is better suited to identifying people with neutral attitudes than more traditional scales that often explicitly exclude neutral items. Interestingly, if one factor analyzes Thurstone scaled items, one can obtain phantom factors that mischaracterize the dimensionality of the items because factor analysis assumes linear IOCs (Spector & Brannick, 2010). Does the type of scaling model implied by the Thurstone IOC map onto your research? Is it a scaling approach you might consider?

Another influential scaling theory in attitude measurement was proposed by Rensis Likert (1932) and it assumes linear tracelines, per Figure 1.1. It is called the *method of summated ratings*. Ironically, Likert's pioneering work on this approach has been overshadowed by the common use of the term "Likert scale" to refer to all kinds of rating scale formats, many of which Likert had nothing to do with. The term "Likert scale" is often a misnomer. The method of summated ratings uses items that are either quite positive or quite negative. Endorsement of each item is usually measured on a five point disagree-agree metric (strongly disagree, moderately disagree, neither, moderately agree, strongly agree). The working assumption is that the more positive a person's attitude towards the attitude object, the more likely he or she will endorse positive items and not endorse negative items; the more negative a person's attitude the toward the attitude object, the more likely he or she will endorse negative items and not endorse positive items. Note that this assumption is quite different from that of Thurstone scaling. Neutral items are explicitly excluded from Likert scaling because they will not elicit the desired IOC. As noted, the typical test of a linear IOC is the item total correlation. The overall attitude is defined as the sum of the scores across the final items, with appropriate reverse coding. Traditional factor analysis also assumes linear trace lines.

Although it is seldom recognized, the traceline used to construct a scale can have implications for behavioral prediction. Just as an item on a scale has a scale value associated with it, so too can a behavioral outcome be conceptualized as such. Using the logic of Thurstone scaling, an individual with a neutral score on a introversion-extraversion scale should be most likely to perform social behaviors that are neutral on the introversion-extroversion dimension; an individual with a moderate degree of extroversion should be most likely to perform behaviors that are moderately extroverted; an individual with a moderate degree of introversion should be most likely to perform behaviors that are moderately introverted. The more discrepant an individual's introversion-extroversion is from the scale value of the behavior, in either direction, the less likely the individual should be to perform the behavior. For a Guttman scale, If an individual's degree of extroversion is less than the degree of extroversion implied by the behavior, then the probability of performing the behavior is zero. However, if the

individual's extroversion matches or exceeds the degree of extroversion implied by the behavior, the probability of performance is 1.0, per Figure 1.2. Note that in both the Thurstone and Guttman cases, statistics other than correlations are needed to capture adequately the relationship between scale scores and behavior.

In sum, as you evaluate existing scales or think about forming your own multi-item scale, you need to think about the type of IOC you want to apply. You should devise a theory of IOCs that is reasonable given your broader theory and research goals. We have outlined three examples of IOCs (linear, step-shaped, ideal point) but you might think of other IOC forms that are better suited to your research purposes. When you devise a scale for purposes of predicting behavior, you may want to match the IOC to the way you believe scale scores relate to behavior, i.e., in a step-shaped, ideal point, or linear fashion. Theory and measurement are intimately intertwined.

ITEM RESPONSE THEORY

Item Response Theory (IRT) is an increasingly popular scaling approach that draws upon the notions of scale values and IOCs described above. It is very much tied to the fundamental concepts of Guttman and Thurstone scaling, although this is rarely acknowledged. It uses different terminology, referring to scale values as *item difficulties* (or difficulty levels) and to IOCs as *item characteristic curves* (ICCs) or *item response functions* (IRFs). People's true scores on the underlying dimension are referred to as *theta* (θ). (Half the battle of understanding IRT relative to traditional scaling theory is orienting to the new nomenclature introduced by IRT). Early versions of IRT focused on dichotomous responses to items (agree-disagree, true-false, pass-fail), but IRT later was expanded to more than two response categories. We introduce it using the dichotomous case because it is easiest to explain. We also retain the terminology of more traditional scaling theory to make it easier for you to integrate IRT concepts with our earlier discussion, with the exception of using the term theta to refer to true scores on the underlying construct dimension (because it is more compact). However, when you read about IRT, you will need to transition to its jargon.

In one variation of IRT, the IOC linking theta to the probability of endorsement of an item takes the form of a (cumulative) logit function. As such, it uses yet a different IOC than Guttman, Thurstone, and Likert scaling. In statistics, when we have a dichotomous outcome (item response) and a continuous predictor (theta), it is common to analyze the data using logistic regression. This is the model form assumed by classic IRT; the dichotomous response to the item is conceptually "regressed" onto the continuous true scores using a logistic model. Figure 1.4 presents an example IOC for the IRT logistic model where the item's scale value is 0. Another IOC sometimes used in

IRT scaling is a cumulative normal probability distribution, which in the IRT literature is called a *normal-ogive* model (which we do not consider here).

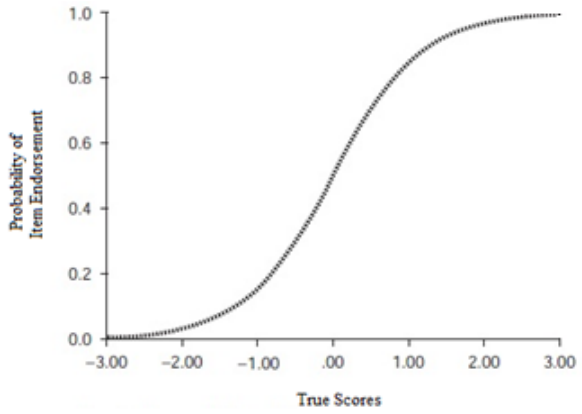

FIGURE 1.4. Logit IOC

Item scale values are defined in IRT as the value on theta where the proportion of people endorsing or "passing" the item is 0.50 or greater. This occurs for the item in Figure 1.4 at a theta value of zero. This is evident if we extend a dotted line rightward from the 0.50 probability point on the Y axis. At the point where it and the curve intersect, we extend a dotted line downwards and see that it intersects with a theta value of 0, per Figure 1.5.

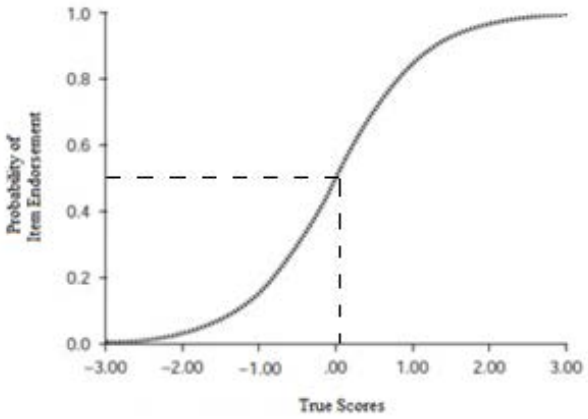

FIGURE 1.5. Scale Value of an Item

This particular scaling model assumes that items with low scale values will be "passed" or "endorsed" by most everyone but items with high scale values will not (in the spirit of a Guttman scale). As an example, suppose we examine the construct of depression in a clinical population and the "items" are depressive symptoms of varying degrees of severity. Most everyone will indicate they are experiencing the mild symptoms but only those with high levels of depression will indicate they are experiencing the severe symptoms. Items that reflect higher symptom severity will show a similar response curve to that of Figure 1.4 but the curve will be shifted to the right on the theta dimension. Items that reflect "mild" symptoms will be shifted to the left of the curve in Figure 1.4. Figure 1.6 presents an example of an item with a scale value of 0 and an item with a scale value of 1.0. We added dashed lines for the latter item so you can see the basis for its scale value. IRT allows us to derive a scale value (or difficulty level) for each candidate item based on the presumed scaling model and the proportion of people who "endorse" or "pass" the item. Another term in the IRT literature for a scale value or difficulty level is an item's *threshold* or *location*.

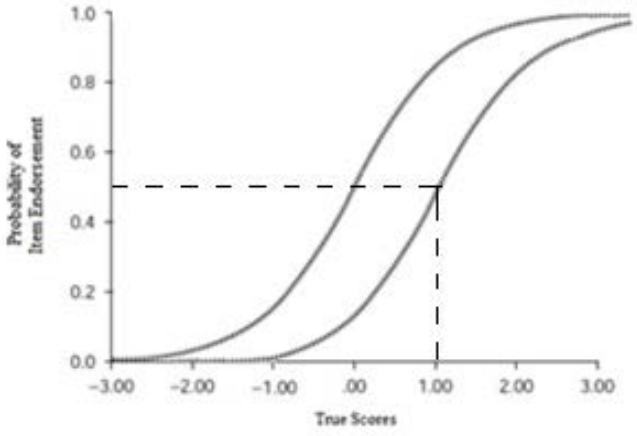

FIGURE 1.6. Two Items with Different Scale Values

IRT analyses often are used to identify item bias for different groups using a technique called *differential item functioning* (DIF). DIF analysis focuses on identifying scenarios where the scale value or difficulty level of an item varies for subgroups. Ideally, for a test to be appropriate for general use, its scale value will *not* vary by subgroup. Items exhibiting DIF might be excluded from the final scale (Edelen, McCaffrey, Marshall & Jaycox, 2009).

Another key concept in IRT scaling is the discriminatory power of an item, or more simply, its *discrimination*. This refers to the ability of an item to discriminate people along the underlying dimension. Using our logistic regression analogy, the discrimination of an item is analogous to the magnitude of the logistic coefficient for a predictor, with larger coefficients being indicative of greater impact on the outcome, everything else being equal. When constructing a scale using IRT, we seek items that have large discrimination and we eliminate items with low discrimination. .

For traditional scales, a total score is based on summing responses to items (e.g., Likert's summated ratings) or by calculating the average scale value of endorsed items (e.g., Thurstone scaling). In IRT, the person's overall score is obtained using a complex maximum likelihood scoring method based on the correspondence between the person's response pattern across items with theoretically derived item scale values. The person is assigned an overall score that has the maximum probability of producing the individual's response pattern across items.

IRT is an elegant theory that offers diverse approaches to scale construction. The theory is too complex for extended summarization here, but like the scaling theories of Guttman, Thurstone, and Likert, it works with the core concepts of scale values, IOCs, item screening, and true score estimation. For good introductory treatments of it, see Baker (2001) and Baker and Seock (2017). For a practical introduction to IRT model types, see de Ayala (2008). For a more advanced treatment, see Raykov and Marcoulides (2018).

CONJOINT AND FUNCTIONAL MEASUREMENT

Conjoint measurement is a scaling theory that is typically used for the analysis of social judgment and decision-making. A consumer psychologist might want to determine the contributions of different types of information about a product on people's evaluations or preferences for that product. In a typical "conjoint task," consumers might be shown 8 different descriptions of a product (e.g., vacuum cleaners) and asked to rank order the descriptions from the one they feel most favorable about through the one they feel least favorable about. The 8 descriptions are strategically structured combinations of the product attributes the researcher is interested in, say, cost of the product (high versus low), performance (high versus low) and ease of use (high versus low). The descriptions typically represent a 2X2X2 factorial combination of the target attributes. Each description appears on a printed card and individuals sort the 8 cards in terms of their preferences. For example, one description might be for a product that is high in performance, high in cost, and low in convenience. Another description might be for a product that is low in performance, low in cost, and high in convenience. Of interest is

how consumers trade off or weight price versus performance, price versus convenience, and performance versus convenience information when forming product evaluations or preferences.

Conjoint analysis has been applied to many substantive areas and contexts. Examples include how social workers make judgments of child abuse based on different case information, how clinicians make judgments of depression severity based on information about patient symptoms, how forest managers evaluate the promise or potential effectiveness of different fire control plans for national parks, how livestock judges at state fairs evaluate the quality of livestock, and how men evaluate possible male contraceptive pills designed to prevent pregnancy. The prototypical experimental design of this research is that people rank order different profiles (hereafter referred to using the generic term *objects*) defined on the basis of a factorial combination of different types of information. The people who do the ranking are often called *judges*.

The end product of a conjoint analysis is what is called a *part worth utility* for each piece of information. The part worth utility is scaled on an interval level metric and reflects the "scale value" of each piece of information on the underlying judgment dimension, such as a preference judgment, a severity judgment, an effectiveness judgment or whatever judgment people are asked to make when rank ordering the profiles. The goal of conjoint measurement is to estimate this part worth utility on an interval-level metric but, remarkably, the judgment task it uses is ordinal in character, namely a rank ordering of objects/profiles. As well, conjoint measurement seeks to order the objects/profiles themselves on the underlying judgment dimension but, again, using an interval level-metric. To us, this is quite a feat, namely, to derive interval level metrics from ordinal level responses. It took creative theorizing to accomplish these goals.

Conjoint analysis has its roots in a measurement framework known as *representational measurement*. Representational measurement uses mathematical representations to characterize empirical relationships between objects, much like the natural sciences. Shortly after the start of the representational measurement movement in the 1870s, it became clear that measurement approaches for the natural sciences did not import well to the social sciences. Conjoint measurement reflected one effort by psychometricians to adapt representational measurement to the measurement of social-psychological constructs. For an introduction to representational measurement, see Luce and Suppes (2002). In this section, we first discuss the core axioms and theorems that underlie conjoint measurement. Next, we present an example to illustrate the key concepts of conjoint analysis. Finally, we discuss functional measurement (Anderson, 1981, 1982), a scaling method popular in psychology that builds upon conjoint measurement principles.

Conjoint Measurement Axioms and Theorems

Conjoint measurement uses mathematics and psychology to accomplish its goals and is noteworthy for integrating perspectives from these two disciplines. It is expressed mathematically in the form of axioms and theorems. *Axioms* are mathematical statements taken as givens. *Theorems* are derived from axioms based on the logic of mathematics. We characterize here the core axioms of conjoint measurement that allow it to generate indices of part-worth utilities. However, in the interest of pedagogy, we do so intuitively rather than with rigorous mathematics. For the latter, see Krantz et al. (1971). We refer to Y as the judgment dimension of interest (e.g., preference, severity, effectiveness). Y is assumed to be continuous and quantitative. Conjoint measurement assumes that the objects or profiles that judges are asked to rank can be *weakly ordered* along this dimension. Weak ordering means that the objects being judged can be ordered on the dimension of interest but that there also can be ties. Ordering implies the mathematical property of *transitivity*. If object 1 is judged more positive than object 2, and object 2 is judged more positive than object 3, than object 1 should be judged more positive than object 3. When judges rank profiles, sometimes violations of transitivity occur, which undermines a key conjoint axiom. Conjoint measurement software usually flags when this occurs. We use an example the case of two informational dimensions that are thought to impact Y, with three levels of each dimension (also referred to as a *factor*). Thus, we construct profiles based on a two factor, 3X3 factorial design. The factors are labelled generically as A and B. Factor A has the levels a1, a2, and a3; Factor B has the levels b1, b2, and b3. The two factors combine to form 9 different objects or profiles, per Table 1.

Conjoint measurement makes the assumption that the location of an object on the Y dimension is an additive function of the part worth utilities of the pieces of information used to describe it. This is reflected in the cell entries of Table 1 in which the part-worth utilities, U, are summed to represent the overall score on the preference dimension. In analysis of variance (ANOVA) terms, this means there is no interaction effect in the way the different pieces of information, or more technically, their utilities, combine to influence Y. It is for this reason that conjoint analysis is often referred to as *additive conjoint measurement*. Extensions of the method to more complex types of functions exist, but our focus here is on the additive model. The axiom/assumption of additivity is controversial because research often finds that information combines to influence Y in non-additive ways (see Anderson, 1981, 1982). In this case, we might use functional measurement (discussed below) to explore different information integration functions.

Table 1: Factorial Design with Utilities

	b1	b2	b3
a1	$U_{a1} + U_{b1}$	$U_{a1} + U_{b2}$	$U_{a1} + U_{b3}$
a2	$U_{a2} + U_{b1}$	$U_{a2} + U_{b2}$	$U_{a2} + U_{b3}$
a3	$U_{a3} + U_{b1}$	$U_{a3} + U_{b2}$	$U_{a3} + U_{b3}$

One set of axioms for conjoint measurement focus on the concept of *cancellation*. These axioms state that if certain combinations of A and B are ordered in a certain way, then other combinations must be ordered in certain ways as well. For example, consider the case of what is called s*ingle cancellation*. If one object is described by information a1 and b1 and another object is described by a1 and b2, we might find the following preference relationship:

$$U_{a1} + U_{b1} > U_{a1} + U_{b2}$$

It follows from this inequality that the utility for b1 must be larger than the utility for b2 because the utility for a1 is present or "held constant" for each object on the two sides of the inequality. If this is the case, then for any instance where another piece of information is held constant in conjunction with b1 and b2, the same relationship should hold. For example,

$$U_{a2} + U_{b1} > U_{a2} + U_{b2}$$

$$U_{a3} + U_{b1} > U_{a3} + U_{b2}$$

A form of cancellation called *double cancellation* also is an axiom of conjoint measurement. Specifically, if

$$U_{a1} + U_{b2} > U_{a2} + U_{b1} \qquad [A1]$$

and

$$U_{a2} + U_{b3} > U_{a3} + U_{b2} \qquad [A2]$$

then this implies that

$$U_{a1} + U_{b3} > U_{a3} + U_{b1} \qquad [A3]$$

This axiom is a bit obscure, but the underlying logic goes something like this: If we add

the left hand sides of A1 and A2 together as well as the right hand sides, this yields

$U_{a1} + U_{b2} + U_{a2} + U_{b3} > U_{a2} + U_{b1} + U_{a3} + U_{b2}$

because a2 and b2 are common to both sides of the equality, we can drop them, yielding

$U_{a1} + U_{b3} > U_{a3} + U_{b1}$

which is inequality A3. Double cancellation is important because it implies many constraints that must exist that can be taken advantage of for the estimation of the part worth utilities (see Michell, 2014).

Yet another axiom of conjoint measurement is known as *solvability*. This condition focuses on variables A and B in the abstract, not necessarily the particular levels of the factors that are used to define the objects in the particular experimental implementation. Solvability requires that variables A and B be complex enough to produce any value of Y and that any change in one factor (increase or decrease) can be compensated for by a change (increase or decrease) in the second factor.

A final axiom is known as the *Archimedean condition*. It states that no utility associated with a given piece of information is either infinitely greater than or infinitely smaller than the utility associated with any other piece of information.

These axioms coupled with other well-known properties of additivity (see Michell, 2014) were used by Luce and Tukey (1964; see also Krantz et al., 1971) to derive theorems to yield measures that (a) preserve the order of preferences among the target objects, (b) provide an interval level metric of the degree of difference in preference between the objects, and (c) provide indices of part worth utilities of the individual pieces of information that combine independently and additively to yield a preference index on Y. The theorem is complex and we do not describe it here. Interested readers can consult Luce and Tukey (1964) and Krantz et al. (1971).

Conjoint measurement has both strengths and weaknesses that we elaborate below. It represents a creative and elegant scaling theory that captures the spirit of representational measurement. It has been widely used in marketing and consumer psychology for product development and it or its variants have the potential for widespread application in a variety of social science domains.

An Example to Illustrate Core Concepts

We adapt an example by Green and Rao (1971) to illustrate the core concepts of conjoint analysis. A consumer is asked to indicate his or her preferences for 8 different types of web-site ads all touting the virtues of the same product. For example, the web sites varied

in the text they used, their use of static images, floating banners, flash videos, pop-ups, and so on. The 8 types of web sites were crossed with five different types of background music, yielding an 8X5 factorial design. The different levels of ad type are indicated by a1 through a8 and the different levels of music type by b1 through b5. The consumer rank ordered the 40 ads in terms of preference for the ad. The researcher seeks to scale the part-worth utility of each ad type and the part worth utility of each type of background music to help structure future marketing strategies.

To evaluate how well conjoint analysis captures true part worth utilities, Green and Rao created a set of part-worth utilities for a hypothetical consumer for each piece of information, as shown in Table 2. The cell entries are the overall preference scores for each ad/music type combination on the underlying preference dimension and are simply the sum of the two applicable part worths for a given cell. A table of the rank values assigned to the 40 ads by the consumer appears in Table 3. The least preferred ad is ranked 1 and the most preferred ad is ranked 40. The ranks were derived from the cell entries in Table 2, so the ranks map perfectly onto the true underlying preference dimension. (This will not always be the case in practice). If the conjoint framework is logically sound, one would expect it to adequately recover the true part worth utilities (signified by the U) for the ad type and for the music type in Table 2 (or, more technically, recover a linear transformation of them). As well, the preferences in the cell entries in Table 2 should be recovered, again, in the form of a linear transformation of those entries. The data were analyzed using Kruskal's (1965) MONANOVA algorithm for conjoint analysis. A stress index yielded by the program that indicates badness of fit was near zero, which is expected given the data perfectly map onto the core assumptions of conjoint measurement. Lack of transitivity, for example, would contribute to a non-zero stress value.

Table 2: True Utilities and True Preferences

	b1 (U=2)	b2 (U=6)	b3 (U=11)	b4 (U=22)	b5 (U=24)
a1 (U=1)	3	7	14	23	25
a2 (U=4)	6	10	17	26	28
a3 (U=6)	8	12	19	28	30
a4 (U=9)	11	15	22	31	33
a5 (U=12)	14	18	25	34	36
a6 (U=18)	20	24	31	40	42
a7 (U=21)	23	27	34	43	45
a8 (U=25)	27	31	38	47	49

Table 3: Consumer Rank Orders

	b1	b2	b3	b4	b5
a1	1	3	8.5	16.5	19.5
a2	2	5	11	21	24.5
a3	4	7	13	24.5	26
a4	6	10	15	28	30
a5	8.5	12	19.5	31.5	33
a6	14	18	28	25	36
a7	16.5	22.5	31.5	27	38
a8	22.5	28	34	39	40

Figure 1.6 presents a scatterplot of the derived vs. actual part worth utilities for the ad types and for music types. Figure 1.7 presents a scatterplot between the cell values generated by the program and the input ranks of Table 2. The analysis generated good interval level estimates of the part worth utilities for both the ad types and music types and also captured well the input rank orders based on the predicted (interval-level) preferences.

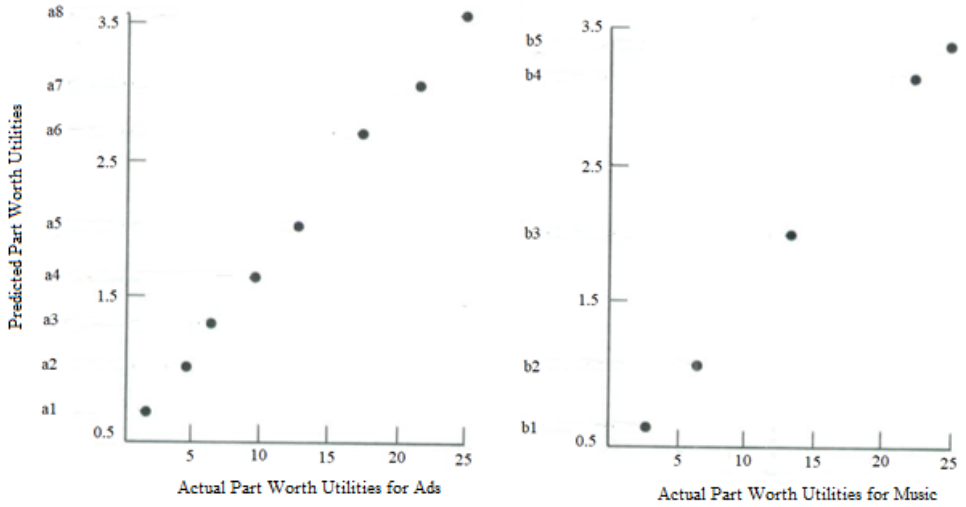

FIGURE 1.6. Predicted and Actual Part Worth Utilities (based on Green & Rao, 1971)

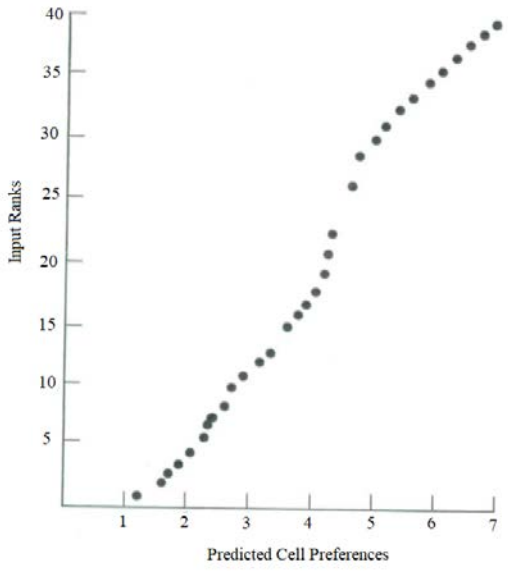

FIGURE 1.7. Input Ranks by Predicted Cell Preferences (based on Green & Rao, 1971)

Note that although the actual part worth utilities ranged from 1 to 25, the predicted part worth utilities generated by the MONANOVA program ranged from 0 to 4. The predicted part worth utilities are not the actual part worth utilities but rather a linear function of them with an arbitrary origin. Because part worth utilities have an arbitrary scale, some researchers transform them to be centered about 0 (thus yielding both negative and positive values) or to range from 0 (lowest utility) to 100 (highest utility). Whatever the case, the predicted utilities are scaled on an interval level metric as are the overall utilities for the 40 cells, which are simply the sum of the component utilities for a given cell.

If a study is conducted on, say, 100 individuals, conjoint analysis can be used to derive part-worth utilities for each individual for each piece of information and then these part worth utilities can be subjected to statistical analyses to evaluate the variability in them across individuals, the central tendency of them across individuals, group differences in them, and correlates of them, either in the form of predictors or outcomes. As well, the overall utility/preference value for a given cell of the design can be subjected to across individual analyses.

In addition to model estimates, most conjoint software performs checks on the input rank data to determine if there are violations of selected conjoint assumptions (e.g., transitivity). The presence of many violations suggest conjoint analysis should be abandoned. Software for conjoint analysis can be expensive because it is popular in

marketing. The freeware R offers packages and analytics that can perform many forms of conjoint analysis.

The Current State of Conjoint Analysis

From the roots described above, conjoint analysis has seen a dizzying number of innovations and extensions (Agarwal, DeSarbo, Malhotra & Rao, 2015), some of which have ventured far from the representational measurement framework in which conjoint measurement is grounded. These innovations include (a) dealing with the presence of non-additivity, (b) specifying methods to deal with large numbers of factors, (c) developing diverse strategies for presenting object information (e.g., pictorially, verbally), (d) developing alternative methods of statistical analysis, and (e) tying the results of conjoint analysis to meaningful behaviors and outcomes in real world settings, among others. It is not our intent to address these issues here. Rather, we merely seek to highlight conjoint analysis as an example of creative scaling theory construction that yields interval level metrics from ordinal level data on an idiographic level in ways that can provide insights into fundamental social judgment phenomena. To be sure, the method has challenges, but from a theory construction standpoint, it has many elegant qualities and reflects the creativity that can be brought to measurement oriented theory construction.

Functional Measurement

One criticism of additive conjoint measurement is that it cannot accommodate well a range of information integration rules other than additivity that empirics suggest are common (Anderson, 1981). Another criticism is that it does not include a formal error theory that recognizes the fact that measures of part-worth utilities and preferences are subject to measurement error. Anderson (1981, 1982) developed a measurement approach closely related to conjoint measurement, called *functional measurement*, that also focuses on social judgments, that is idiographic in character, that makes use of factorial designs in the definition of objects/profiles to be judged, that seeks to isolate part worth utilities (but they are called *scale values* and are further parameterized by the introduction of importance weights for them), and that yields interval level metrics for the scale values and the preference judgments for objects/profiles on the underlying preference dimension. However, unlike conjoint measurement, functional measurement incorporates an error theory and also can accommodate a more diverse set of information integration rules.

A primary difference between functional and conjoint measurement is that

individuals directly rate each object/profile on a rating scale, such as those discussed in Chapter 14 in the main text. Anderson presumes that by using sound psychometric practices, researchers can usually obtain ratings that reasonably approximate interval level properties; and he offers diagnostics and correctives for when this may not be the case. Anderson also formulates and implements an error theory by having individuals repeat the judgment process on multiple occasions, in the spirit of test-retest reliability designs. The multi-trial data are then analyzed using analysis of variance based procedures to yield error adjusted parameter estimates. For details on the theoretical foundations and measurement model of functional measurement, see the definitive two volume set by Anderson (1981, 1982).

MULTIDIMENSIONAL SCALING

The final creative approach to scaling that we consider is *multidimensional scaling* (MDS). All of us are familiar with maps that allow us to locate points of interest on a two dimensional coordinate system of north-south and east-west. The thesis of multidimensional scaling is that people have mental maps of "objects" (e.g., people, events, places) that locate those objects in a psychological coordinate system. The scaling task is to construct representations of these mental maps and specify the coordinates of objects on them. MDS reflects a creative scaling theory by leveraging an analogy between physical maps and mental maps.

MDS works with distance or proximity scores between objects. Typically, the distance or proximity scores are ratings of similarity or dissimilarity between objects. Just as we can re-construct a city-based map of the United States by knowing the pairwise distances between major cities like Baltimore, Chicago, Indianapolis, Los Angeles, Miami, New York and New Orleans, so too can we construct a mental map for an individual based on the perceived similarity (called *proximities*) or dissimilarity (called *distances*) of different objects. For example, suppose we ask rheumatologists to review 20 different patient profiles and to rate how similar one profile is to another profile for all possible pairs of profiles. The result would be a 20X20 symmetric matrix of proximities between the patients. Or, a consumer psychologist might ask consumers to rate how similar 15 different breakfast cereals are to one another on a pairwise basis. The result would be a 15X15 symmetric matrix of proximities. The proximity/distance matrices are subjected to MDS analysis to identify the underlying coordinate system and the respective coordinates. One can perform MDS in two dimensional space (as with geographic maps), in three dimensional space, in four dimensional space or in k dimensional space. With three dimensions, graphical representations of MDS results are challenging and with four or more dimensions, specialized visual tools are required.

The statistical algorithms for MDS are complex, so we do not describe them here. Interested readers are referred to Borg and Groenen (2010) and Schiffman, Reynolds and Young (1981). Suffice it to say that they can be applied to ratio, interval and ordinal level data and do a reasonably good job of recovering population structures of known dimensional structure if the proximity measures are valid and reliable. For example, Schiffman et al. (1981) calculated the pairwise distances between 10 major cities in the United States and the analyzed these distances using MDS algorithms. A two coordinate system emerged with a plot that appears in Figure 1.8 using the estimated coordinates of each city on the two dimensions. The analysis captured well the true geographic city locations.

FIGURE 1.8. MDS of Inter-City Flying Distances (from Schiffman et al., 1981)

MDS works best when proximities are measured with ratio or interval level metrics, or reasonably close approximations to them. Nevertheless, MDS methods also are available for ordinal data. Judgments of similarity can be obtained for each pair of objects using rating scales (e.g., with anchors “exactly the same” to “completely different, with appropriate adverb qualifiers in between and using metrics that provide good precision, per Chapter 14). Alternatively, sorting tasks can be used in which individuals sort objects into piles based on their similarities. Given that all possible pairs of objects need to be judged, prudence is required about the number of objects studied. For example, the number of pairs for 10 objects is 45, for 15 objects it is 105, for 20 objects it is 190, and for 25 objects it is 300. If the number of pairs is large, sometimes researchers will have

participants split up the rating tasks over multiple sessions or with breaks between the rating tasks. Often the judgments are made on several similarity ratings scales so that the judgments can be averaged, thereby increasing their reliability through the cancellation of random error. The order of presentation of pairs of stimuli usually is randomized.

MDS analysis yields indices of goodness/badness of fit for a given dimensional model in terms of its ability to reproduce the input distance ratings by individuals based on that model. If a two dimensional model fits poorly, then the investigator might test if a three dimensional model fits well. If a three dimensional model also provides poor fit, then a four dimensional model might be tested. And so on. A key issue in MDS is the determination of the number of dimensions needed to adequately model the distances between objects. This is usually accomplished by comparing goodness of fit statistics for solutions with differing numbers of dimensions. Parsimony and interpretability of the solution are also taken into account.

The most common index of badness of fit for a model is called a *stress index* and it ranges from 0 to 1.00. The lower the value, the better the model fit, everything else being equal. The index reflects the average absolute disparity (using root mean square averages) between the model predicted and observed distances/proximities, adjusted by a scale factor to force it to range from 0 to 1.0. Once a satisfactory model is isolated, the coordinates for each object on each dimension are of interest.

MDS can be applied to idiographic (individual level) or nomothetic (group level) data. When group level data are analyzed, the group mean or median distance/proximity score is calculated for each pair of objects and then subjected to MDS analysis. We discuss later how to work with individual differences in MDS frameworks.

Interpreting Dimensions and Coordinates

One task MDS analysts face is to assign meaning to the solution. There are three general strategies that are common. The first strategy is to examine how objects align themselves on a given dimension in terms of their locations on them and then infer meaning based on those locations. For example, Kruskal and Wish (1978) report an example where individuals rated the similarity of 12 different countries, with the resulting group-level solution consisting of two dimensions. The results are shown in Figure 1.9a. Sometimes, interpretation is clarified by rotating the dimensions to alter the coordinate values but in a way that preserves the configural relationships among all of the objects. This is shown in Figure 1.9b. with the rotated solution characterized by the dashed lines. It illustrates that there is nothing sacred about the coordinate values per se on the dimensions, only the configural relationships that they imply. In this instance, Kruskal and Wish characterized the two dimensions as (1) underdeveloped versus developed, and (2) pro-western versus

pro-communist. These represent the major dimensions upon which countries are perceived or classified by the individuals studied, i.e., they constitute key elements of people's mental maps for countries.

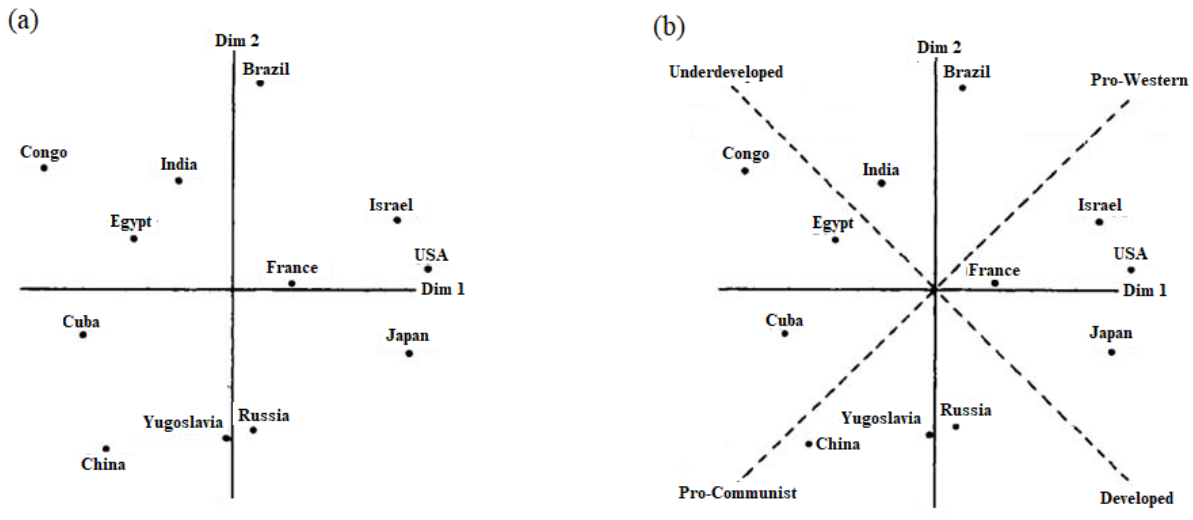

FIGURE 1.9. Dimension Rotation (based on Kruskal & Wish, 1978) al., 1981)

The second method uses the coordinate values yielded by the MDS in conjunction with supplemental ratings for each object. This approach was used in a study by Robinson and Bennet (1995) who conducted an MDS of 46 deviant workplace behaviors presented in the form of scenarios.[1] A two dimensional solution emerged. Prior to the MDS, Robinson and Bennet had expert judges rate on five point scales each of the 46 behaviors on six attributes, namely whether the behavior was (1) unintentional/intentional, (2) not serious/serious, (3) not harmful to company/harmful to company, (4) not harmful to individuals/harmful to individuals, (5) very unethical/ethical, and (6) covert/overt. The mean rating for the judges for each behavior on each attribute was calculated. They then regressed each attribute score onto the two coordinate values from the MDS for each of the two dimensions across the 46 behaviors (hence the N for the regression analysis was 46 and there were two predictors, the coordinate value for dimension 1 and the coordinate value for dimension 2). For the first dimension, there

[1] Respondents were asked to complete only a subset of the 990 possible pairwise judgments to make the task manageable. However, Robinson and Bennet ensured that all possible combinations were rated by at least 20 respondents. The mean distance score for the 990 pairwise behaviors were used as input for the MDS analysis.

were sizeable and statistically significant, positively signed regression coefficients with the following attributes: not serious/serious, not harmful to company/harmful to company, and not harmful to individuals/harmful to individuals. Thus, one end of the dimension reflected deviant behaviors that tended to be not serious, not harmful to the company, and not harmful to the targeted individuals, while the other end reflected deviant behaviors that were serious, harmful to the company, and harmful to the targeted individuals. The dimension was therefore labeled "minor versus serious deviance."

For dimension 2, the regression coefficients that were sizeable and statistically significant included the attributes of not harmful to company/harmful to company (positive association), not harmful to individuals/harmful to individuals (negative association), and covert/overt (negative association). Thus, one end of this dimension reflected behaviors that were harmful to individuals, not harmful to the organization, and covert, while the other end of the dimension reflected behaviors that were harmful to the organization, not harmful to individuals, and overt. Robinson and Bennet labeled the second dimension "interpersonal versus organizational deviance." Thus, their research suggests that people tend to think of deviant behaviors along two dimensions, minor versus serious deviance, and interpersonal versus organizational deviance.

A third strategy sometimes used to interpret MDS results is known as *nearest neighbor analysis*. This takes many forms but the idea is not to focus so much on the naming of dimensions but instead to focus on objects that cluster together and are close to one another in the dimensional space. Robinson and Bennet called attention to four notable deviant behavior “neighborhoods.” One closely grouped set of behaviors was labeled “production deviant” and included leaving early, taking excessive breaks, intentionally working slow and wasting resources. Another closely grouped set of behaviors was labeled “property deviance” and included sabotaging equipment, accepting kickbacks, lying about hours worked, and stealing from the company. A third closely grouped set of behaviors was labeled “political deviance” and included showing favoritism, gossiping about co-workers, blaming co-workers, and competing non-beneficially. The final set of neighbor-like behaviors was labeled “personal aggression” and included sexual harassment, verbal abuse, stealing from co-workers and endangering coworkers. Production deviant behaviors were in the minor-organizational quadrant of the two dimensional space, property deviance behaviors were in the serious-organizational quadrant, political deviance behaviors were in the minor-interpersonal quadrant, and personal aggression was in the serious-interpersonal quadrant.

Uses of Multidimensional Scaling

MDS can be used for either exploratory or confirmatory purposes. For example, theories

of emotions differ in the hypothesized number and nature of emotion dimensions thought to underlie emotion recognition. MDS can be used to test these theories in a confirmatory way. By contrast, exploratory MDS does not have strong theory about the dimensional structure/content of mental maps and approaches research from the perspective of discovering qualities of people's mental maps, in the spirit of Chapter 11 in the main text. MDS has been applied to a wide range of interesting phenomena in this regard. Examples include the study of perceptions of political protest behaviors, perceptions of crimes, perceptions of nation-states, perceptions of neighborhoods, color perceptions, Morse code confusions, perceptions of faces, perceptions of political candidates, perceptions of kinship groups, perceptions of market structures in public markets, perceptions of languages, perceptions of patients, perceptions of physicians, perceptions of consumer products, perceptions of treatments for cancer, and perceptions of food, among others.

MDS analyses extend beyond the mere identification of mental maps. One also can use MDS identified dimensions to identify correlates of those dimensions. Falbo (1977), for example, used MDS to build a cognitive map of power strategies used by college students who wrote essays on the topic "How I Get My Way." The essays were content analyzed and 16 core strategies were identified. Expert judges then rated the similarity of the 16 strategies, and MDS was performed on the mean judgments made by the judges. The analysis yielded a two dimension solution. The first dimension was labeled rational/nonrational, with exemplar rational strategies being use of reason, compromise, and appeals to expertise; exemplar non-rational strategies included emotion manipulation, deceit, and evasion. The second dimension was labeled direct versus indirect, with direct strategies represented by such strategies as persistence, simple statements, and assertion and indirect strategies represented by strategies such as hinting and thought manipulation. Analytic methods were used that then related the coordinate values for the strategies on each dimension to a personality scale measuring Machiavellianism and to positive or negative peer evaluations. Falbo found that the use of rational power strategies was associated with positive peer evaluations and that Machiavellianism was associated with the use of indirect and nonrational strategies. .

Another interesting use of MDS is for product development in business and marketing. Based on an MDS analysis, one might decide to develop a new product with certain featured attributes. The decision to position a new product in a particular perceptual space identified by MDS might be critical to its success. MDS also can give insights into the key attributes consumers use to classify or group products by virtue of the dimensions that emerge in the analysis. Nearest neighbor analysis for the proposed positioning of a new product also can be useful. Note that MDS modeling does not rely on experimenter-imposed rating dimensions that might bias classification schemes; the

similarity judgments made by consumers are based on their internalized psychological maps free of experimenter imposed judgment dimensions. This is a strength of MDS more generally. Contrary to what some analysts assert, MDS cannot easily identify completely novel products because it is constrained by the domain of objects it is applied to (Schiffman et al., 1981). Having said that, MDS can identify product classes that contain few competitor products and that may be an easier class to break into.

MDS and Individual Differences

All of the above examples used group-level data in the form of mean proximity or distance scores. It is possible to apply MDS at the level of individuals and then aggregate or summarize the data based on individual profiles. Indeed, the conclusions one might draw based on aggregate means can be quite different than those that are more respective of individual differences. You can conduct traditional MDS for a single individual with any of the standard MDS software, but to aggregate or classify the individual profiles in a statistically rigorous way, a program known as INDSCAL should be used. Input into INDSCAL is a separate matrix of distance scores for each individual, one stacked on top of the other. INDSCAL then seeks a common group space across individuals as well as a weight space for the 10 different individuals from which parameters from their individual solution can be derived. You obtain a common stress index across individuals as well as a separate stress index for each individual. The subject weights provide insights into the weight that each individual places on a given dimension from the common solution. For statistical details, see Borg and Groenen (2010), Kruskal and Wish (1978) and Schiffman et al. (1981).

We have only scratched the surface of MDS and its potential. Again, our goal was not to describe the underlying theory and method in depth but to highlight the basic ideas and to give you a sense of the creative theorizing that entered into its development.

CONCLUDING COMMENTS

Theory construction at the level of measurement can address different topics. Chapter 13 in the main text addressed theory construction about sources of random error and sources of systematic error in the measurement process. Factors relevant to these sources of error can vary as a function of the population studied, the testing context, the construct being measured, and the timing of measurement. A theory construction mindset that invokes these facets can help you strengthen the measurement protocols you use for your research *and* will benefit others as you publish empirical results surrounding your theories of random and systematic error. Chapter 14 in the main text encouraged you to construct

theories focused on question comprehension, mental judgments made in response to questions, and response translation of those judgments into response formats provided by researchers. Each of these processes also can vary as a function of the population studied, the testing context, the construct being measured, and the timing of measurement. Again, a theory construction mindset for these matters can strengthen the measurement protocols you use and will benefit others as you publish research surrounding those theories.

In this primer, we focused on a different facet of measurement theory construction, namely how observed measures map onto the qualities and properties of the underlying construct and the metric qualities of those measures relative to that construct. We considered three types of scaling theories, (a) scaling theories for multi-item scales, (b) conjoint and functional measurement theory, and (c) multidimensional scaling theory. For multi-item scales, we discussed the potential for theory construction surrounding different forms of item tracelines and the implications for interpreting overall scores on the scale. Conjoint measurement and multidimensional scaling were used as examples of highly creative theory construction as focused on the scaling process per se. The scientists who evolved the notions of the different forms of tracelines, who derived conjoint/functional measurement, and who invented multidimensional scaling were all extremely creative measurement theorists. Our challenge to you is to bring such creativity to devise your own measurement theories.

REFERENCES

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.

Andrews, J., Hops, H., Ary, D. et al. (1991). The construction, validation and use of a Guttman scale of adolescent substance use: An investigation of family relationships. *Journal of Drug Issues*, 21, 557-572.

Agarwal, J., DeSarbo, W., Malhotra, N. & Rao, V.R. (2015). An interdisciplinary review of research in conjoint analysis: Recent developments and directions for future research. *Customer Needs and Solutions*, 2, 19–40.

Baker, F. (2001). The basics of item response theory. New York: ERIC Clearinghouse on Assessment and Evaluation.

Baker, F, & Seock, H. (2017). The basics of item response theory using R. New York: Springer.

Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156.

Borg, I. & Groenen, P. (2010). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

de Ayala, R. J. (2008). *The theory and practice of item response theory*. New York: Guilford.

Drasgow, F. L., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, 3, 465–476.

Edelen, M.O., McCaffrey, D.F., Marshall, G.N. & Jaycox, L.H.(2009). Measurement of teen dating violence attitudes: an item response theory evaluation of differential item functioning according to gender. *Journal of Interpersonal Violence*, 24, 1243-1263.

Edwards, A. L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.

Edwards, A. & Gonzalez, R. (1993). Simplified successive intervals scaling. *Applied Psychological Measurement*, 17, 21-27.

Falbo, T. (1977). Multidimensional scaling of power strategies. *Journal of Personality and Social Psychology*, 35, 537-547.

Green, B.F. (1954). Attitude measurement. In G. Lindzey (Ed.). *Handbook of social psychology*, Vol 1, pp. 335-369. Reading, Mass: Addison-Wesley.

Green, P. E. & Rao, V.R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, 8, 355-363.

Guttman, L.A. (1944). A basis for scaling qualitative data. *American Sociological Review*, 91, 139–150.

Jaccard, J. & Blanton, H. (2005). he origins and structure of behavior: Conceptualizing behavior in attitude research. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change*. Mahwah, NJ: Erlbaum.

Kay, P. (1964). A Guttman scale model of Tahitian consumer behavior. Southwestern *Journal of Anthropology*, 20, 160-167.

King, C. & Christensen, A.(1983). The relationship events scale: A Guttman scaling of progress in courtship. *Journal of Marriage and Family*, 45, 671-678.

Kline, R. (2016). *Principles and practice of structural equation modeling*. New York: Guilford.

Krantz, D. H., Luce, R., Suppes, P & Tversky, A. (1971). *Foundations of measurement, Vol. I: 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*. Thousand Oaks, CA: Sage.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1–55.

Lindemann, D.F. (2003). A Guttman scale for assessing condom use skills among college students. *AIDS and Behavior*, 7, 23-27.

Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: a new scale type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.

Luce, R. D. & Suppes, P. (2002). Representational measurement theory. In H. Pashler and Wixted, J. (Eds.). *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology*. pp. 1–41, New York: Wiley.

Mead, A.D., & Meade, A.W. (2010). Item selection using CTT and IRT with unrepresentative samples. Paper presented at the twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA.

Michell, J. (2014). *An introduction to the logic of psychological measurement.* New York: Taylor and Francis.

Raykov, T. & Marcoulides, G. (2018). *A course in item response theory and modeling with Stata*. College Station, Texas: Stata Press.

Robinson, S. & Bennett, R. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *The Academy of Management Journal*, 38, 555-572.

Rosow, I., & Breslau, N. (1966). A Guttman health scale for the aged. *Journal of Gerontology*, 21, 556-559.

Schiffman, S., Reynolds, M. & Young, F. (1981). Introduction to multidimensional scaling: Theory, method, and applications. New York: Academic Press.

Spector, P. & Brannick, M. (2010). If Thurstone was right, what happens when we factor analyze Likert scales? *Industrial and Organizational Psychology*, 3, 502–503.

Stoeber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, 17, 222-232.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-554.