1

Constructing Multi-Item Scales: An Introduction to 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 3 of the main text, is an example of a scaling theory, albeit a crude one. This primer introduces traditional multi-item scaling theories used in social science research. To be sure, there are a wide assortment of scaling theories and I can't possibly cover all of them here. My emphasis is on scaling theories that you might use in developing a multi-item scale for use in an RET because you are unhappy with existing instruments or a satisfactory measure does not exist.

THE CONSTRUCTION OF MULTI-ITEM SCALES

In this section, I first focus on the general practice of item screening from a more mundane practical level and then address theory-based item screening using the concept of item operating characteristics. I then introduce the general logic of Item Response Theory (IRT).

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.

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 3, 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 3 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. I find most people are understanding if I am transparent with them about my purposes. I 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 I 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, I either eliminate items that have base rate problems (i.e., show highly skewed response patterns) or I 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 3. 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 for a scale 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), Nießen et al. (2019) 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. I discuss these in Chapter XX.

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. I discuss methods for determining sample sizes necessary to achieve desired MOEs in the main text and provide a program for doing so on the program tab of this website (see the program 'MOE for Regression'). For example, for an expected reliability of 0.85 and a MOE of 0.15, you will need a sample size of approximately 60 to 65.

Finally, once ill-behaved items have been screened or revised, it is useful to subject the remaining items to cognitive response testing per Chapter 3. 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 I 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. I 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.

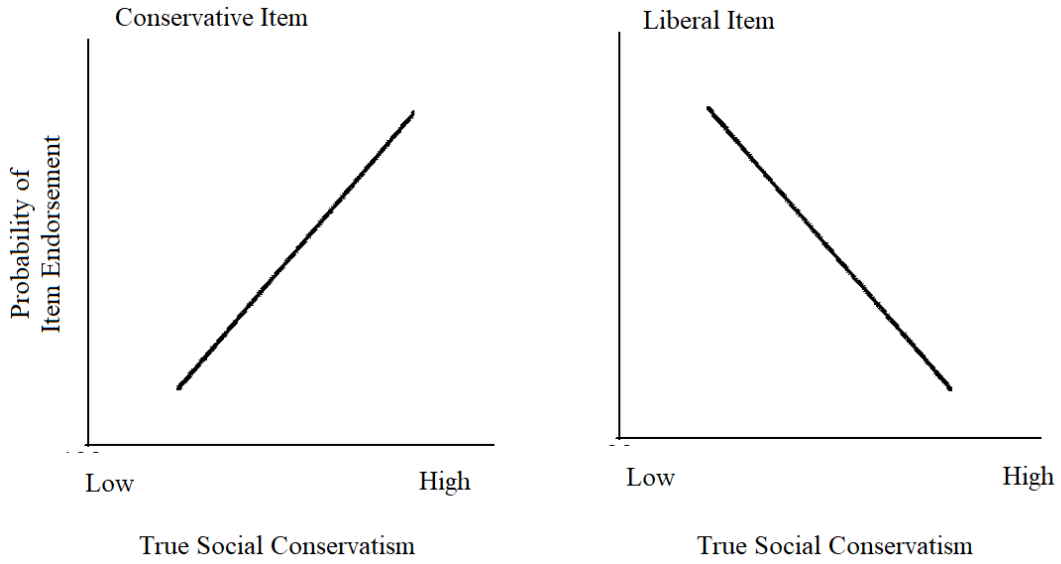

FIGURE 1.1. Linear Traceline

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, I 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.

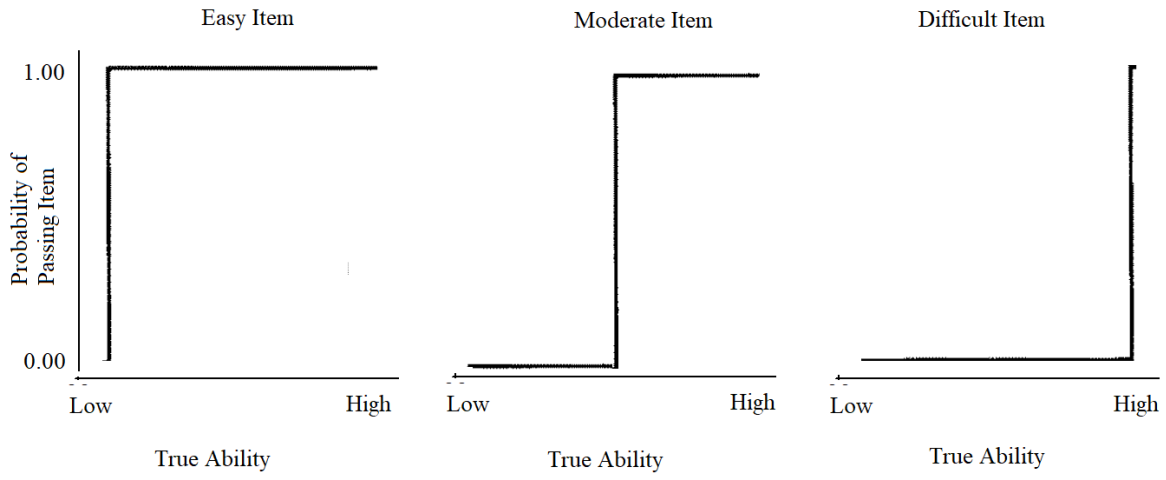

FIGURE 1.2. Step-Shaped Traceline

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 a scale value of -4.5, 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.

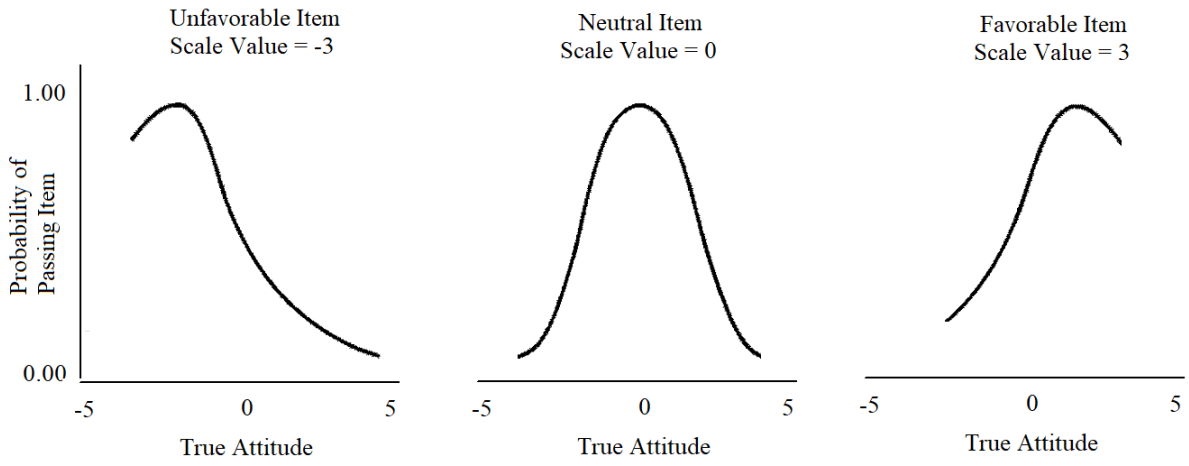

FIGURE 1.3. Ideal Point Traceline

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).

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 different from that of Thurstone scaling. Neutral items are explicitly excluded from Likert scaling because they do 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 tracelines.

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 an 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 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 or that was applied. You should devise an approach to IOCs that is reasonable given your broader theory and research goals. I 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 Screening using Measurement Invariance

Once you have reduced the number of items on your scale by ensuring they have an appropriate IOC, another round of item screening should be enacted to remove items that

show blatant measurement non-invariance. I discuss measurement non-invariance in Chapter 3 and provide a primer for it on my webpage in the Resource section for chapter 3.

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. I introduce it using the dichotomous case because it is easiest to explain. I 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 I do not consider here).

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 I extend a dotted line rightward from the 0.50 probability point on the Y axis. At the point where it and the curve intersect, I extend a dotted line downwards so that it intersects with a theta value of 0, per Figure 1.5.

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 I 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. I 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*.

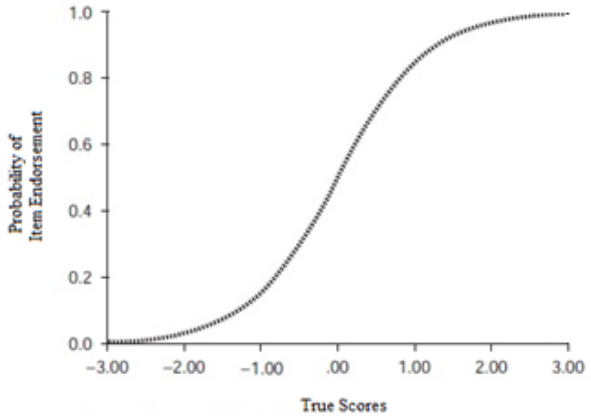

FIGURE 1.4. Logit IOC

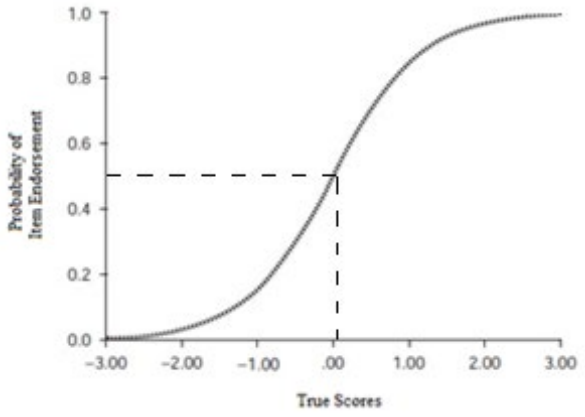

FIGURE 1.5. Scale Value of an Item

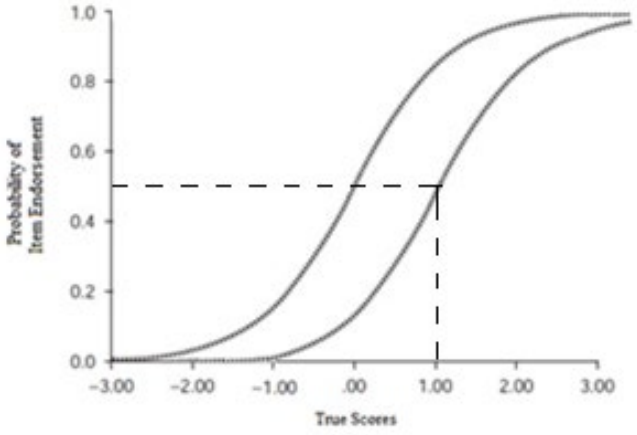

FIGURE 1.6. Two Items with Different Scale Values

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).

CONCLUDING COMMENTS

Chapter 3 in the main text described some sources of random and 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. When you approach measurement, I urge you to adopt a mindset that invokes these facets to help you strengthen the measurement protocols you use. Chapter 3 also emphasized a conceptual framework that encouraged you to think about the three processes of 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 mindset in which you want to tailor the framing and wording of your questions as a function of these facets is helpful.

REFERENCES

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.

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.

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.

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.

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

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

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.

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.

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.

Nießen, D., Partsch, M., Kemper, C. & Rammstedt, B. (2019). An English-language adaptation of the social desirability–gamma short scale (KSE-G). *Measurement Instruments for the Social Sciences*, 2, 2.

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

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

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.