1 WHAT IS RATING SCALE ANALYSIS?

The central topic of this volume is *rating scale functioning*. Rating scale functioning refers to the degree to which ordinal rating scales with three or more categories, such as Likert-type rating scales used in attitude surveys, can be interpreted and used in a psychometrically sound way. Researchers who are concerned with rating scale functioning evaluate their rating scale data for properties such as those listed in Table 1.1.

First, rating scale functioning is concerned with *rating scale category ordering*. When rating scale categories are functioning well, higher categories in a rating scale should reflect higher levels of the construct being measured. For example, in a scale designed to measure empathy, participants who *Strongly Agree* with statements asking whether they exhibit empathetic behaviors should have higher levels of empathy than participants who *Agree* with those statements. Next, *rating scale category precision* refers to the degree to which individual rating scale categories make meaningful distinctions between participants with respect to the construct. When rating scales function well, each category reflects a unique level of the construct. For example, there should be a meaningful difference in the level of empathy between participants who *Strongly Agree* and those who *Agree* with survey items. Finally, *rating scale category comparability* analyses help researchers investigate whether rating scale categories have a similar interpretation across assessment components or subgroups of participants. For example, the

Table 1.1 Indices Used in Rating Scale Analysis

Rating Scale Properties	*Guiding Question for Rating Scale Analysis*
Rating scale category ordering	To what extent do higher rating scale categories indicate higher locations on the construct?
Rating scale category precision	To what extent do individual rating scale categories reflect distinct ranges on the construct?
Rating scale category comparability	To what extent do rating scale categories have a similar interpretation and use across assessment components or subgroups of participants?

difference in the level of the empathy required to *Strongly Agree* and *Agree* should be similar across participants with different levels of education. There are many analytic techniques through which researchers can explore rating scale functioning. However, methods based on Rasch models and item response theory (IRT) models are particularly suited to this approach. In this book, we explore methods for examining rating scale functioning using these methods.

The purpose of this book is to provide readers with an overview of rating scale analysis, choices involved in rating scale analysis, and practical guidance on how to conduct such analyses with their own survey data. The analyses are based on Rasch models and IRT models, with some references to classical test theory to highlight the advantages of the Rasch and IRT approaches.

The organizing principle for this book is that rating scale functioning must be examined each time a survey is administered before inferences can be made from participant responses. Currently, most of the information about rating scale analysis is contained in a substantially longer and technically sophisticated book (Wright & Masters, 1982) or in a few select chapters in longer books (e.g., Engelhard & Wind, 2018) and articles (Linacre, 2002; Wind, 2014) whose target audience is methodologists in the area of psychometrics. The current volume is targeted to a wider audience, and readers need only basic training in psychometrics and familiarity with Rasch and IRT approaches in order to understand and apply all of the concepts. Readers who are beginners in psychometrics can focus on the interpretation and practical use of rating scale analysis methods.

What Is Item Response Theory?

IRT is a paradigm for the development, analysis, and evaluation of measurement instruments (e.g., surveys or tests) for latent variables (i.e., constructs), such as attitudes, abilities, or achievement levels in the social sciences. IRT is sometimes referred to as latent trait theory, strong true score theory, or modern measurement theory, among other names (Embretson & Reise, 2000). IRT is a broad framework in which numerous models are available for different kinds of measurement tools and assessment contexts. Each model is characterized by assumptions that are reflected in model parameters and formulations. These differences have implications for the information that each model provides about participants, items, and latent variables. For example, many IRT

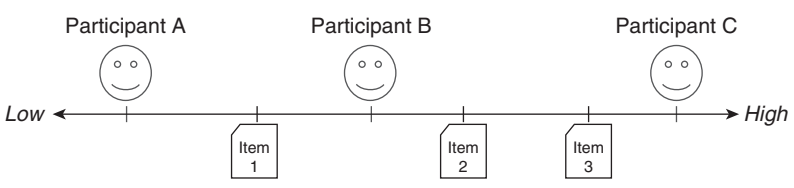
models assume or require approximate *unidimensionality*. In the context of IRT, unidimensionality means that a single latent variable can be used to explain most of the variation in the item responses of interest.

The basic idea underlying unidimensional IRT models is that latent variables can be expressed using a single linear continuum on which participants and items have unique locations.

For example, in an educational assessment designed to measure student achievement in physical science, students whose understanding of physical science concepts are relatively advanced would have higher locations on the construct compared to students whose understanding of physical science concepts are less advanced. Likewise, items that require more proficiency in physical science to produce a correct response (i.e., difficult items) would have higher locations on the construct compared to items that require less proficiency in physical science to produce a correct response (i.e., easy items). As another example, in a measure of depression, participants with more severe depressive symptoms would have higher locations on the construct compared to participants with less severe depressive symptoms. Items that reflect more frequent or severe depressive symptoms would have higher locations compared to items that reflect less severe symptoms.

Figure 1.1 illustrates the concept of a latent variable expressed as a linear continuum. The horizontal double-ended arrow represents the latent variable (e.g., depression) as a unidimensional continuum ranging from low (e.g., low levels of depression) to high (e.g., high levels of depression). Three participants are shown above the continuum and three items are shown below the continuum, with locations indicated using vertical lines. Participant A has the lowest location (e.g., lowest level of depression), followed by Participant B, followed by Participant C, who has the highest location (e.g., highest level of depression). Likewise, Item 1 has the lowest location (this item requires minimal

Figure 1.1 Illustration of Participant and Item Locations on a Latent Variable



depression for positive responses), followed by Item 2, followed by Item 3, which has the highest location (this item requires severe depression for positive responses).

IRT models describe the probability for a certain type of response (e.g., an accurate response to a multiple-choice item in an educational assessment or a rating of "*Strongly Agree*" to an item in an attitude survey) as a function of the difference between participant locations and item locations on this linear continuum. For example, in Figure 1.1, Participant C would be expected to respond positively to all three items because Participant C's location exceeds those of the items. In contrast, Participant B would be expected to respond positively to Item 1, but not Item 2 or Item 3, whose locations exceed that of Participant B. Participant A would not be expected to respond positively to any of the items.

When there is evidence of acceptable fit between the responses and the model assumptions or requirements (i.e., good model-data fit), the participant and item location estimates can be interpreted as location estimates on a common linear scale that represents the latent variable. This common metric for participants and items is a major advantage of IRT beyond scaling techniques that focus on the decomposition of variance in number-correct or average scores, such as classical test theory (CTT; Crocker & Algina, 1986; Gulliksen, 1950).

IRT for Rating Scale Data

A popular use of IRT is to analyze data from measurement instruments that include rating scales, such as attitude surveys with Likert-type response scales (Likert, 1932). There are a number of IRT models that can be usefully applied to multicategory (i.e., polytomous) data, including the Rating Scale Model (Andrich, 1978), the Partial Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1997), the Graded Response Model (Samejima, 1969, 1997), and models from Mokken Scale Analysis (Mokken, 1971); each of these models will be discussed in turn later in this volume. When they are applied to rating scale data, these models provide information with which researchers can evaluate psychometric properties, compare participant and item locations on the construct, and use the results to inform the revision, interpretation, and use of measurement instruments for research and practice. Importantly, IRT models also provide a variety of tools that can be used to conduct rating scale analysis.

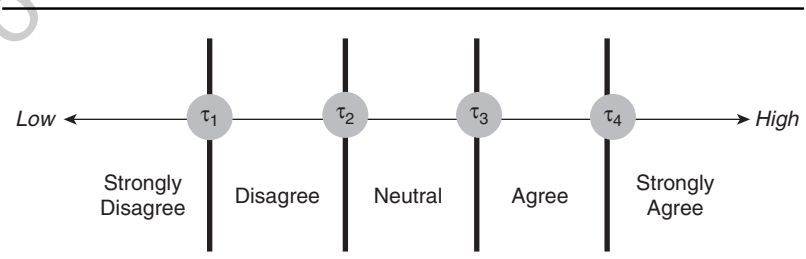
What Is Rating Scale Analysis?

Rating scale analysis is a procedure for evaluating rating scale functioning in item response data that includes participant responses in three or more ordered categories (e.g., Likert-type responses) for evidence of useful psychometric properties at the level of individual rating scale categories. To illustrate the concept of rating scale functioning, Figure 1.2 illustrates a typical Likert-type rating scale with five ordered categories. This figure has a similar interpretation to Figure 1.1, but it illustrates the locations of *rating scale categories* on a latent variable. The horizontal double-ended arrow represents a latent variable (e.g., empathy) as a unidimensional continuum ranging from low (e.g., low levels of empathy) to high (e.g., high levels of empathy). Below the arrow, five ordered rating scale categories are shown, ranging from *Strongly Disagree* to *Strongly Agree*. Thick, evenly spaced vertical lines show the four transition points ("thresholds"; τ_k) that correspond to the five adjacent rating scale categories. The location of each threshold on the latent variable is marked with a circle. The thresholds are monotonically nondecreasing as the latent variable progresses from low to high (i.e., $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$), such that higher levels of the latent variable (e.g., more empathy) would be required to provide a rating in higher categories.

How Is Rating Scale Analysis Different From Other Survey Analyses?

Survey researchers frequently evaluate their data for evidence of validity, reliability, and fairness. These investigations often draw on methods such as factor analysis, internal consistency statistics

Figure 1.2 Illustration of Evenly Spaced, Monotonic, Ordinal Rating Scale Categories



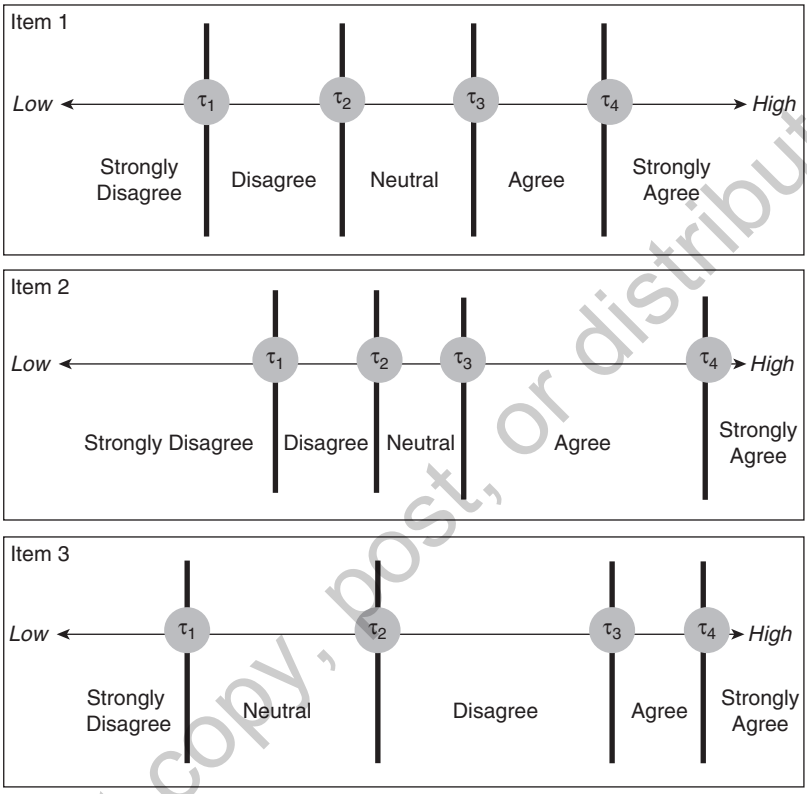
(e.g., alpha), and IRT analyses. Most routine survey analyses focus on *overall item difficulty estimates,* which assume that rating scales can be interpreted as they are illustrated in Figure 1.2. When researchers use these analyses, they assume that the ordinal categories in a rating scale are evenly spaced with respect to the latent variable (e.g., the difference in the level of empathy required to respond in *Strongly Agree* and *Agree* is the same for responses to Item 1 and Item 2) and that increasing categories reflect increasing levels of the latent variable (e.g., responses in the *Neutral* category always reflect higher levels of empathy than responses in the *Disagree* category). If these properties hold, researchers can make meaningful comparisons between participant responses to different items, and across subgroups of participants (e.g., gender subgroups) with regard to the interpretation of the scale categories. Moreover, evidence of acceptable rating scale functioning provides support for interpreting overall item-level analyses and total score analyses related to validity, reliability, and fairness.

Rather than assuming the properties illustrated in Figure 1.2 hold in all situations, rating scale analysis uses indices from IRT models to empirically explore the structure of ordinal rating scales. The purpose of such analyses is to ensure a meaningful interpretation of responses within and across components of a scale and subgroups of participants. For example, rating scale analyses could help researchers identify a scenario such as the one illustrated in Figure 1.3 (discussed next).

Figure 1.3 shows the rating scale structure for three items from a hypothetical survey. For Item 1, rating scale analysis may reveal that the categories are evenly spaced and monotonically nondecreasing as the latent variable progresses from low to high. However, for Item 2, *Strongly Disagree* and *Agree* represent wider ranges of the latent variable compared to *Disagree*, *Neutral*, and *Strongly Agree.* In addition, the relative distance between category thresholds is not consistent across the scale. For Item 3, the category spacing is different still from that of Item 1 and Item 2. Moreover, the category *ordering* is also different from that of the first two items: For Item 3, responding in the *Disagree* category requires higher levels of the latent variable (e.g., more empathy) than responding in the *Neutral* category, such that $\tau_3 < \tau_2$, whereas the opposite order is true for Item 1 and Item 2. Suffice it to say, it would be difficult to justify interpreting participant responses to Item 3 in the same way as their responses to Item 1 and Item 2.

Rather than focusing on total scores (as in classical test theory) or on overall item difficulty or discrimination parameters (as in many polytomous IRT analyses), rating scale analysis focuses on examining the degree

Figure 1.3 Illustration of Different Rating Scale Structures for Individual Items



to which the *individual categories* in an ordinal rating scale have a meaningful interpretation that is consistent across elements of the assessment procedure, such as across items, persons, or subgroups of items or persons. This information is important because evidence of acceptable rating scale functioning ensures meaningful interpretation of the directionality of rating scales, informs the interpretation of differences between categories, and ensures a comparable interpretation of categories between items, components of a scale, or subgroups of items or persons. In summary, rating scale analysis supplements routine survey analyses related to validity, reliability, and fairness to help survey researchers ensure that their rating scales have meaningful interpretations.

What Are the Requirements for Rating Scale Analysis?

Rating scale analyses can be conducted on item responses across a range of survey research applications designed to measure unidimensional constructs using ordinal rating scales.[1] There are no universal requirements for the maximum number of rating scale categories, minimum sample sizes for persons or items, or the proportion of missing data that always make a dataset eligible for rating scale analysis. However, each of these issues is important and should be considered before conducting rating scale analysis. This section provides guidance on practical considerations related to each of these topics.

Number of Rating Scale Categories

Rating scale analysis can be conducted when rating scales include three or more ordered categories (Wright & Masters, 1982). Examination of survey research literature reveals that researchers tend to disagree about the appropriate maximum length for collecting useful data from survey instruments in general (e.g., Borgers et al., 2004; DeCastellarnau, 2018; Krosnick et al., 2002; Linacre, 2002; Weijters et al., 2010). Similarly, there is no universal maximum value for the number of categories that is appropriate for conducting rating scale analysis. However, there are some general guidelines that researchers can use to determine whether their scale length may be appropriate for rating scale analysis.

Technically, there is no upper limit to the number of scale categories that could be included in rating scale analyses, and researchers could conduct the analyses included in this book on instruments with many rating scale categories, such as Everett's (2013) Social and Economic Conservatism Scale, whose rating scale categories range from 0 to 100. However, there are several practical challenges with long scales. First, if a scale includes many categories, it may be unlikely that there will be enough responses in each category to meaningfully interpret rating scale analysis indicators. This challenge is particularly relevant with small samples, where there may not be sufficient variation within a sample to identify participants at each level of the scale. For example, as will be

[1]Rating scale analysis has not yet been considered in the context of multidimensional IRT. However, analysts can apply the techniques illustrated throughout this book to their multidimensional measures by analyzing item responses for each construct separately.

discussed in Chapter 3, one potential cause for disordered categories is a relatively low frequency of observations within a category. In addition, the volume of information that rating scale analysis provides is directly related to the number of categories. Interpreting and using information about rating scale ordering, precision, and comparability for many categories across multiple items may not be practical in all survey research contexts.

The perspective taken in this book is that the number of categories that are suitable for rating scale analysis varies across contexts. Researchers should rely on guidance from the literature, practical experience with their population(s) of interest, and empirical evidence from rating scale analysis methods to determine how many categories are appropriate for their rating scales and subsequent rating scale analyses. The techniques presented in this book provide practical tools with which researchers can empirically evaluate the effectiveness and contribution of each category to a measurement procedure. As we will see throughout the book, rating scale analysis helps researchers use empirical evidence to determine whether more or fewer rating scale categories may be needed in a survey research context and whether this number should be fixed or can vary across items. In addition, rating scale analysis provides empirical evidence to support the use or omission of neutral categories. These issues are considered throughout the volume, and Chapter 6 provides a specific discussion on decisions related to combining or omitting categories based on the results from rating scale analysis.

Participant Sample Size

Sample size requirements for rating scale analysis are directly tied to the modeling choices that researchers make and their plans for statistical tests that they will conduct using scores from the measurement instrument. Among the models considered in this book, those based on nonparametric IRT (see Chapter 5) are the least stringent in sample size requirements (some researchers have reported these analyses with as few as 30 participants), while those based on IRT models with item discrimination parameters, such as the Generalized Partial Credit Model (see Chapter 4), require the largest samples (usually at least 100 or more participants). Methods based on Rasch models (see Chapters 2 and 3) require moderate sample sizes that are generally attainable in most survey research settings (although larger samples are preferred, samples of around 50 participants are sufficient for many purposes, see Linacre, 1994).

Item Sample Size

Minimum sample sizes for items in rating scale analysis also reflect item sample size requirements for the models used to conduct those analyses. For Rasch models (see Chapters 2 and 3), item and person sample size requirements are symmetric: The precision of information about people depends on the number of items (and the number of categories), and the precision of information about items depends on the number of people. Because measurement models are not statistical significance tests, it is not possible to conduct a traditional power analysis (Cohen, 1969) to identify an exact minimum sample size for the number of items in an instrument for rating scale analysis. Similar to recommendations for the maximum number of categories in a rating scale, it is not straightforward to provide a critical value or equation with which to identify the minimum number of items. In my experience conducting rating scale analysis on scales in a variety of disciplinary areas, including counseling (Cook et al., 2021), empathy (Wind et al., 2018), language learning (Wind et al., 2019), learning motivation (Wang & Wind, 2020), music performance (Wesolowski et al., 2016), and teacher evaluation (Wind, Tsai, et al., 2018), I have found that scales consisting of around ten or more items tend to exhibit strong overall psychometric properties (e.g., high values of reliability coefficients) while also providing a manageable amount of information about rating scale functioning. Reflecting this experience, the example used in this volume is a version of the Center for Epidemiological Studies Depression (CES-D) scale (discussed later in this chapter) that consists of ten items. I recommend that researchers who have shorter instruments still conduct rating scale analysis; the techniques included in this book will provide useful information about rating scale functioning that can be interpreted in the same way as with longer instruments. However, analysts should keep in mind that few items may negatively impact the overall reliability of the instrument and precision of information that can be gleaned about participants. Likewise, scales with many items can also pose psychometric challenges and impact data quality due to issues such as participant fatigue.

Missing Data

Researchers who conduct surveys often encounter missing data (Bodner, 2006), which occur for various reasons (Little & Rubin, 2002). The choice of model for rating scale analysis determines how missing data can be handled. Among the models included in this book, those based on Rasch measurement theory (see Chapter 2) are particularly

amenable to missing data. Specifically, Rasch models can be applied when data are missing as long as there are common observations with which to "link" participants across items and to "link" items across participants (Schumacker, 1999). For example, if participants respond to items in common with other participants, Rasch models can provide location estimates for participants who have not responded to all of the items in the instrument. The same is true for items: Rasch models can provide item location estimates even when some participants have not responded to the item. In contrast, researchers who use non-Rasch IRT models (see Chapter 4) and nonparametric IRT models (see Chapter 5) for rating scale analysis typically impute new values for missing responses before applying the model (Hagedoorn et al., 2018; van der Ark & Sijtsma, 2005).[2]

Relatedly, rating scale analyses should only be conducted on meaningful item responses. For example, some survey instruments allow participants to indicate that an item is not applicable or that they have no opinion about the content in an item. Such response options usually cannot be meaningfully placed within the ordinal rating scale. As a result, they should be treated as missing data and not included in the analysis. Along the same lines, response patterns that reflect careless responding should be handled following best practices in survey research (Arias et al., 2020; Goldammer et al., 2020) and generally should not be included in rating scale analyses.

How Should Researchers Select a Model for Rating Scale Analysis?

This book presents rating scale analysis methods based on several IRT models. Each of the models discussed in this book offers valuable information that can help researchers evaluate their survey instruments from the perspective of rating scale analysis. However, the models have some important differences in the types of data that can be modeled, how they reflect different overall modeling goals, and how they reflect goals specific to rating scale analysis. Table 1.2 provides an overview of these characteristics for the models used in this book. We will refer to

[2]There has been some preliminary research on using nonparametric IRT models with missing data that does not require imputation methods (Wind, 2020; Wind & Patil, 2016), but these techniques have not been considered in the context of rating scale analysis.

Table 1.2 Overview of Models for Rating Scale Analysis

		Rasch Models (Chapters 2 and 3)				*Parametric Non-Rasch Models (Chapter 4)*		*Nonparametric Item Response Models (Chapter 5)*
				Many-Facet Rasch (MFR) Model				
Considerations for Survey Research		*Rating Scale Model*	*Partial Credit Model*	*Rating Scale Many-Facet Rasch Model*	*Partial Credit Many-Facet Rasch Model*	*Generalized Partial Credit Model*	*Graded Response Model*	*Mokken Scale Analysis*
Types of Data	Response scale includes three or more ordered categories	X	X	X	X	X	X	X
	Response scale varies across items		X		X	X	X	
Overall Modeling Goals	To evaluate survey data for adherence to fundamental measurement requirements (unidimensionality, local independence, invariance)	X	X	X	X			X

(Continued)

	To obtain linear-scale estimates of item, person, and other facet locations	X	X	X		X		X				
	To estimate person and item locations controlling for context-specific explanatory variables			X		X						
	To estimate person and item locations controlling for differences in item slopes							X		X		X
Practical Goals for Rating Scale Analysis	To evaluate rating scale functioning for an entire set of items, without distinguishing between individual items	X		X								
	To evaluate rating scale functioning					X						

(Continued)

Table 1.2 Overview of Models for Rating Scale Analysis *(Continued)*

	Rasch Models (Chapters 2 and 3)				*Parametric Non-Rasch Models (Chapter 4)*		*Nonparametric Item Response Models (Chapter 5)*
			Many-Facet Rasch (MFR) Model				
Considerations for Survey Research	*Rating Scale Model*	*Partial Credit Model*	*Rating Scale Many-Facet Rasch Model*	*Partial Credit Many-Facet Rasch Model*	*Generalized Partial Credit Model*	*Graded Response Model*	*Mokken Scale Analysis*
specific to explanatory variables in an assessment procedure (e.g., person subgroups or item subsets)							
To evaluate rating scale functioning specific to each item		X		X	X	X	X
To evaluate category ordering	X	X	X	X	X		
To evaluate category precision	X	X	X	X	X	X	X

Table 1.2 throughout the book as we consider rating scale analysis techniques based on each model. We revisit the topic of model selection for rating scale analysis in Chapter 6.

What Can Be Learned From Rating Scale Analysis?

Put simply, rating scale analysis helps analysts learn about the quality of their rating scales from a psychometric perspective. Rating scale functioning is an empirical property of item response data that needs to be explored each time a survey is administered. Rather than assuming that rating scale categories are ordered as expected, describe unique ranges of the latent variable, and have comparable interpretations for all items and participants, rating scale analysis helps researchers verify these properties and identify directions for further research or improvement to a measurement instrument. Rating scale analysis supplements other psychometric analyses to help researchers ensure that they can meaningfully interpret the results from participant responses to ordinal rating scales.

What Will This Book Help Researchers Do With Their Data?

This book aims to help researchers identify useful techniques for exploring rating scale functioning in their data and to provide guidance in interpreting the results, along with resources for applying these analyses. Specifically, this book will help researchers use polytomous IRT models to empirically gauge the degree to which participant responses to ordinal rating scales display evidence of psychometrically defensible rating scale functioning. Such information is essential for the meaningful interpretation and use of rating scale data.

Recent developments in statistical software have made rating scale analysis relatively straightforward for analysts who have a basic working knowledge of psychometrics and psychometric software. This book includes online supplemental materials at **https://study.sagepub.com/researchmethods/qass/wind-exploring-rating-scale-functioning** that demonstrate the application of rating scale analysis techniques using R packages (R Core Team, 2021), Winsteps (Linacre, 2016), and Facets (Linacre, 2020). Readers can adapt the provided code for use with their own data.

The remaining chapters are organized as follows. Chapter 2 introduces polytomous models for exploring rating scale functioning based on Rasch measurement theory (Rasch, 1960); these models have several useful properties that make them particularly well suited to rating scale

Table 1.3 CES-D Scale Item Stems

Item Number	*Item Stem*
1	I was bothered by things that usually don't bother me.
2	I did not feel like eating; my appetite was poor.
3	I felt that I could not shake off the blues even with help from my family or friends.
4[a]	I felt I was just as good as other people.
5	I had trouble keeping my mind on what I was doing.
6	I felt depressed.
7	I felt that everything I did was an effort.
8[a]	I felt hopeful about the future.
9	I thought my life had been a failure.
10	I felt fearful.
11	My sleep was restless.
12[a]	I was happy.
13	I talked less than usual.
14	I felt lonely.
15	People were unfriendly.
16[a]	I enjoyed life.
17	I had crying spells.
18	I felt sad.
19	I felt that people dislike me.
20	I could not get "going."

[a]These items are intended to be reverse-coded prior to analysis.

analysis. Chapter 3 continues the discussion of Rasch models for rating scale analysis by demonstrating the application of these models to explore rating scale functioning, along with the interpretation of the results. Chapter 3 includes step-by-step examples and illustrations of rating scale analysis techniques that are supplemented with online resources for applying the analyses using statistical software. Chapter 4 provides a theoretical overview of several popular non-Rasch IRT models that can be used to explore rating scale functioning. Chapter 5 presents a nonparametric approach to rating scale analysis. In Chapter 6, the book concludes with a summary of the topics covered in previous chapters, a discussion of practical choices and considerations for rating scale analysis, and resources for further study.

Introduction to Example Data

To illustrate the application of the methods discussed in this book, example analyses and results will be provided using analyses of data based on an administration of the CES-D scale (Radloff, 1977) as reported by Donny et al. (2015). The CES-D scale is a self-report measure of depression made up of 20 items that ask participants to report the frequency of various symptoms over the previous week using a four-category response scale (1 = *Rarely or none of the time [less than 1 day]*; 2 = *Some or a little of the time [1–2 days]*; 3 = *Occasionally or a moderate amount of time [3–4 days]*; 4 = *Most or all of the time [5–7 days]*). Four of the items require reverse-coding prior to analysis. After recoding, scores range from 0 to 60, with lower scores indicating lower levels of depression and higher scores indicating higher levels of depression. According to the original author of the scale, the stated intended use of the CES-D is as a screening instrument to identify individuals or groups who may be at risk for depression; scores greater than or equal to 16 indicate potential depression (Radloff, 1977). Details about the instrument and a downloadable version of the items are available at http://bit.ly/CES-D_inst. Table 1.3 shows the item stems. For analyses in this book, we use a recoded version of the ratings that range from 0 to 3.

The CES-D scale was selected as the example data for this book for several reasons. First, the CES-D scale was intended to function as a unidimensional measure of a single construct (depression; Radloff, 1977); as a result, it is well suited for analysis with unidimensional IRT models. Second, the CES-D items are in the public domain, and there are several real datasets available with responses to the items online (e.g., data from Donny et al., 2015). Third, this instrument includes a four-category ordinal response scale that is similar to those in many other surveys or questionnaires. Fourth, there have been numerous psychometric evaluations of the CES-D in published research (e.g., Cosco et al., 2020; González et al., 2017; Macêdo et al., 2018), but at the time of this writing, there have not been any published rating scale analyses of this instrument. Together, these characteristics make the CES-D scale a good candidate for an accessible and relevant demonstration of rating scale analysis using an IRT approach.

To facilitate the example analyses in this book, data from a recently published application of the CES-D scale were consulted. Specifically, the CES-D scale data used in the illustrations were collected as part of a larger study related to the impact of reduced-nicotine standards for

cigarettes among individuals who regularly smoked cigarettes (Donny et al., 2015). In this application, no details about the CES-D were reported related to rating scale analysis. The original data included responses from 839 participants who responded to the 20-item CES-D scale. The original data are publicly available for download from the National Institute on Drug Abuse data sharing website link for the Donny et al. study: https://datashare.nida.nih.gov/study/nidacenicp1s1.

For data security purposes, the illustrations and examples are based on data that were simulated using the parameters obtained from an analysis of the CES-D responses at the baseline time point from Donny et al. (2015) with the generalized partial credit model (Muraki, 1997). Readers can download the simulated version of the CES-D data from the online supplement at **https://study.sagepub.com/researchmethods/qass/wind-exploring-rating-scale-functioning** in order to complete the example analyses for this book.

Resources for Further Study

Readers who are new to IRT in general may find the following general introductory IRT texts helpful for learning about this approach in more detail:

DeAyala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Readers who would like to learn more about IRT models for polytomous data may find the following texts helpful:

Nering, M. L., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models.* Routledge.

Ostini, R., & Nering, M. L. (2005). *Polytomous item response theory models* (Vol. 144). Sage.