2 RASCH MODELS FOR RATING SCALE ANALYSIS

This chapter introduces and illustrates two popular measurement models that facilitate rating scale analysis: The Rating Scale model (RSM; Andrich, 1978) and the Partial Credit model (PCM; Masters, 1982), along with the Many-Facet Rasch model (MFRM; Linacre, 1989), which can be specified as an extension of both of these models. These models belong to the family of Rasch measurement theory models (Rasch, 1960; Wright & Mok, 2004), which is a useful framework for rating scale analysis (discussed further below).

Chapter 2 begins with a brief overview of Rasch measurement theory and Rasch models in general. Then, these models are introduced and illustrated using the example CES-D data. Chapter 3 provides a detailed illustration of rating scale analysis using the selected Rasch models.

What Is Rasch Measurement Theory?

Rasch measurement theory (Rasch, 1960) is a theoretical framework based on the premise that principles of measurement from the physical sciences should guide measurement procedures in the social and behavioral sciences. Georg Rasch proposed a theory for social and behavioral measurement that can be summarized in four requirements:

> (1) The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison;
>
> (2) and it should also be independent of which stimuli within the considered class were or might also have been compared.
>
> (3) Symmetrically, a comparison between two individuals should be independent of which particular stimuli with the class considered were instrumental for the comparison;
>
> (4) and it should also be independent of which other individuals were also compared on the same or on some other occasion.
>
> (Rasch, 1961, pp. 331–332)

Together, these four requirements constitute *invariant measurement.* Rasch (1977) used the term *specific objectivity* to describe specific situations in which invariant measurement is approximated. Approximate

adherence to invariant measurement is considered a prerequisite for measurement in the context of Rasch measurement theory. In other words, from the perspective of Rasch measurement theory, it is not appropriate to interpret and use the results from measurement procedures unless there is evidence that the requirements for invariant measurement are appropriately satisfied.

The requirements for invariant measurement are related to two other requirements that characterize Rasch measurement theory. First, Rasch measurement theory requires that item responses adhere to *unidimensionality*. Unidimensionality occurs when one latent variable (i.e., construct) is sufficient to explain most of the variation in item responses. In the context of the CES-D measure of depression mentioned in Chapter 1, unidimensionality would imply that participants' level of depression is the primary variable that determines their responses. Second, Rasch measurement theory requires that item responses reflect *local independence*. Local independence occurs when participant responses to individual items are not statistically related to their responses to other items after controlling for the primary latent variable. In the CES-D scale, local independence implies that participants' responses to one item (e.g., Item 1: *I was bothered by things that usually don't bother me*) do not influence their responses to another item (e.g., Item 2: *I did not feel like eating; my appetite was poor*) beyond what could be predicted given their level of depression. One common cause of violations of local independence in survey research is item stems that contain the same or nearly the same statements. For example, researchers have observed violations of local independence in surveys that contain pairs of statements that are nearly identical but oriented in opposite directions. Using participant responses to the Interpersonal Reactivity Index measure of empathy, Yaghoubi Jami and Wind (2022) observed a violation of local independence between Item 16: *After seeing a play or movie, I have felt as though I were one of the characters* and Item 12, which is a reversed version of nearly the same statement: *Becoming extremely involved in a good book or movie is somewhat rare for me*. We will discuss methods that researchers can use to evaluate unidimensionality and local independence later in this chapter.

What Are Rasch Models?

Rasch models are measurement models that are theoretically and mathematically aligned with Rasch measurement theory (Rasch, 1960, see Chapter 1). Rasch models are mathematically similar to several

other item response theory (IRT) models, such as the one-parameter logistic model (Birnbaum, 1968), but the theoretical perspective underlying the development and use of the model is different. Specifically, Rasch models serve as a guide for evaluating the characteristics of item response data; in other IRT approaches, models are selected whose parameters offer a good *representation* of the characteristics of the data. The major difference between the Rasch approach and other IRT models is that Rasch models use theory to evaluate the characteristics of item responses, whereas other IRT models use the characteristics of item responses to identify and select a model.

The simplest Rasch model is the dichotomous Rasch model for item responses (x) scored in two ordered categories ($x = 0, 1$). These responses are often observed in multiple-choice achievement tests or surveys where participants are asked to agree or disagree with statements. The dichotomous Rasch model states that the probability of Participant n scoring $x = 1$ rather than $x = 0$ on Item i is determined using the difference between the participant's location on the construct (e.g., the participant's level of depression) and the item's location on the construct (e.g., the level of depression required to agree with an item).

The equation for the dichotomous Rasch model appears in the literature in two formats that are mathematically equivalent but describe the model in slightly different terms: The exponent (exp) format, and the log-odds (ln) format. We will start with the log-odds format, which is visually simpler and clearly illustrates key characteristics of the theory underlying Rasch measurement:

$$\ln\left(\frac{P_{ni(x=1)}}{P_{ni(x=0)}}\right) = \theta_n - \delta_i, \tag{2.1}$$

In Equation 2.1, θ_n is the location of Participant n on the construct (i.e., person ability) and δ_i is the location of Item i on the construct (i.e., item difficulty).[1] In words, Equation 2.1 states that the log of the odds that Participant n provides a correct or positive response ($x = 1$), rather than an incorrect or negative response ($x = 0$) on Item i is determined by the difference between the participant location and the item location on the construct. When the difference between person

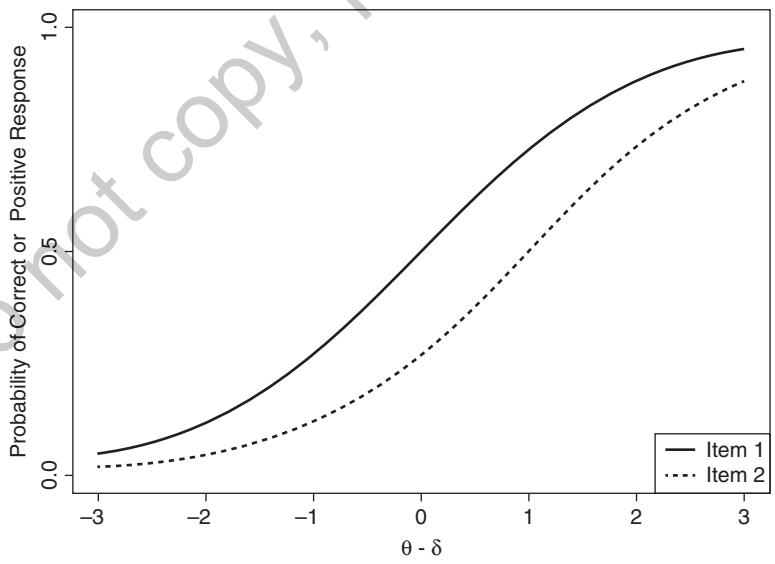
[1]In the original presentation of Rasch measurement theory, Rasch (1960) used the Greek letter "β" to represent person locations. In this text, the Greek letter "θ" is used for alignment with other recent publications on Rasch models and with the non-Rasch models that are presented in Chapter 4.

locations and item locations favors the person, this means that the person is more likely to score 1 than 0. In this case, item difficulty (i.e., the level of the construct required for a correct or positive response) is lower than the person's location on the latent variable (i.e., the person's level of the construct).

For example, in Figure 1.1, Participant B would be expected to provide a correct or positive response to Item 1 because the person location exceeds the item location. When the difference favors the item, this means that the person is more likely to score 0 than 1. In this case, the item difficulty exceeds the person's location. Participant A would be expected to provide an incorrect or negative response to Item 1 because the item location exceeds the person location.

Figure 2.1 illustrates this relationship using an item response function (IRF) for Item 1 and Item 2, both of which were scored in two categories ($x = 0, 1$). In the figure, the x-axis shows the latent variable, expressed as a logit (log-odds) scale. In many Rasch and IRT applications, logit scale estimates range from around –3 to 3 logits that reflect increasing levels of the latent variable as the logit scale progresses

Figure 2.1 Example Item Response Functions for the Dichotomous Rasch Model



from low to high. The y-axis shows the conditional probability for a response in category 1 ($x = 1$). The lines show the expected pattern of response probabilities according to the dichotomous Rasch model for Item 1 (solid line) and Item 2 (dashed line). As participant locations on the construct increase (e.g., higher levels of depression), the probability for a positive rating ($x = 1$) also increases for both items.

The IRFs in Figure 2.1 reflect the requirements for invariant measurement, unidimensionality, and local independence as defined in Chapter 1 and earlier in this chapter because the difference between participant and item locations on the latent variable is sufficient to predict a response in category 1. In addition, for all locations on the x-axis, Item 1 is easier than Item 2, such that item ordering is invariant across participant locations on the latent variable.

It is also possible to state the dichotomous Rasch model equation using an exponent form such that the term on the left side of the equation is the *probability* for a response in category 1 rather than in category 0. This version of the model equation is mathematically equivalent to Equation 2.1, but it is presented differently. Specifically, log odds are transformed to probabilities.

The exponent format of the dichotomous Rasch model highlights the comparison between response categories. In the case of the dichotomous Rasch model, these categories are $x = 0$ and $x = 1$. The exponent format is useful for understanding Rasch models for rating scale data (discussed later in this chapter). The exponent form of the dichotomous Rasch model can be stated as:

$$\frac{P_{ni(x=1)}}{P_{ni(x=0)}} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}. \tag{2.2}$$

In Equation 2.2, the parameters are defined the same way as they were in Equation 2.1.

Polytomous Rasch Models for Rating Scale Analysis

Building on the dichotomous Rasch model, researchers have proposed measurement models for data in three or more ordered categories (i.e., polytomous data), such as data that are obtained from attitude surveys or educational performance assessments. Polytomous Rasch models share the same basic requirements as the dichotomous Rasch model. However, unlike the dichotomous model, a score of $x = 1$ is *not* expected to become increasingly likely with increasing participant

locations on the construct because scores in higher categories (e.g., $x = 2$ and $x = 3$) become more probable as participant locations on the construct increase. In the context of the CES-D scale, as participant depression increases, they are more likely to respond in a higher category.

Figure 2.2 illustrates this relationship for a rating scale with five ordered categories ($x = 0, 1, 2, 3, 4$). The x-axis shows the logit scale that represents the construct, and the y-axis shows the conditional probability for a rating in category k given participant and item locations. Separate lines show the conditional probabilities for each category in the rating scale. Moving from left to right on the x-axis, the probabilities for higher rating scale categories increase while the probabilities for lower rating scale categories decrease. In other words, as participant locations on the construct increase, they are more likely to respond in higher categories. This basic relationship can also be seen in Figure 2.3, which shows a polytomous IRF based on the rating scale category probabilities in Figure 2.2. In Figure 2.3, the y-axis shows the model-expected rating at each location on the logit scale, which is shown along the x-axis. As logit-scale locations increase, expected ratings increase.

Figure 2.2 Rating Scale Category Probability Curves

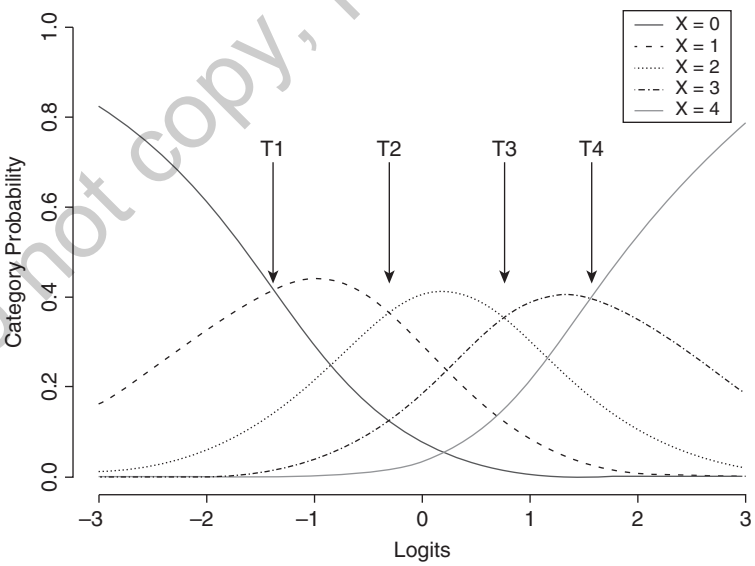
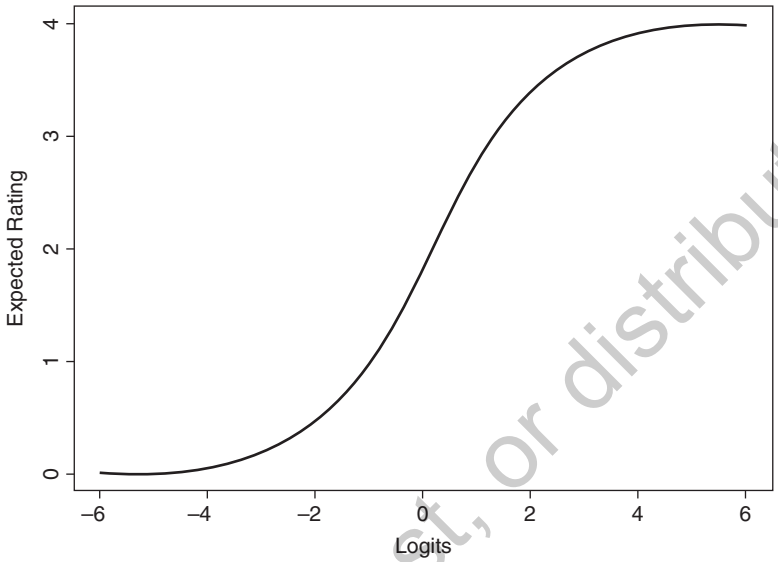


Figure 2.3 Expected Ratings



Why Are Polytomous Rasch Models Useful for Rating Scale Analysis?

Polytomous Rasch models are particularly suited to rating scale analysis for two main reasons. First, they are characterized by the same requirements as the dichotomous Rasch model. As a result, researchers can use polytomous Rasch models to evaluate item responses for evidence that they adhere to fundamental measurement properties. Discrepancies between model requirements and item responses alert researchers to components of a measurement procedure (e.g., items) that warrant revision, aspects of theory about the construct that may warrant reconsideration, and directions for future research.

Second, polytomous Rasch models model the probability for a rating in a given rating scale category using an *adjacent categories probability formulation.* This means that the model is based on comparisons between pairs of categories, as we saw in the exponent form of the dichotomous Rasch model (Equation 2.2), where the probability for $x = 1$ was compared to the probability for $x = 0$. The polytomous Rasch model equation can also be stated in exponent form, and it is

nearly identical to the dichotomous version of the model (Equation 2.2) with two major differences. First, the polytomous Rasch model compares the probability for a response in category k (e.g., *Strongly Agree*) to the probability for a response in category $k - 1$ (e.g., *Agree*). Second, the equation includes a threshold parameter (τ) that represents the difficulty associated with a specific rating scale category. In Figure 2.2, thresholds are the intersection points between adjacent categories. We discuss thresholds in more detail later in this section.

The exponent form of the polytomous Rasch model is written as follows:

$$\frac{P_{ni(x=k)}}{P_{ni(x=k-1)} + P_{ni(x=k)}} = \frac{\exp[\theta_n - (\delta_i + \tau_k)]}{1 + \exp[\theta_n - (\delta_i + \tau_k)]}, \tag{2.3}$$

where θ_n and δ_i are defined as before. In Equation 2.3, the item difficulty parameter (δ) is combined with a *rating scale category threshold parameter* (τ_k). This means that we no longer consider items on their own, but we now consider items in combination with a set of ordered rating scale categories. As it is defined in many Rasch measurement theory applications (Andrich, 1978, 2013), the threshold parameter (τ) is the point on the logit scale at which the probability for a rating in category k is equal to the probability for a rating in category $k - 1$. For example, in the CES-D scale, the first threshold represents the level of depression at which participants are equally likely to respond in category 2 (*Some or a little of the time)* as they are to respond in category 1 (*Rarely or none of the time*).

For a rating scale with k categories, there are $k - 1$ threshold parameters. In Figure 2.2, there are four arrows that correspond to the thresholds between each category in a five-category rating scale. The comparison between adjacent categories (category k rather than category $k - 1$) is an important feature of polytomous Rasch models that distinguishes them from several other IRT models and facilitates analyses that are particularly useful for exploring the structure of rating scales. Specifically, this formulation allows analysts to identify disordered rating scale categories when they occur (e.g., Item 3 in Figure 1.3). Briefly, disordered categories occur when the level of the construct required to respond in each category does not match the intended order. For example, in the context of the CES-D scale, category disordering may occur if higher levels of depression are required to respond in category 4 (*Most or all of the time*) than are required to respond in category 3 (*Occasionally or a moderate amount of time*). We

discuss category disordering further in the remaining chapters of this book. Several other popular polytomous IRT models, such as the Graded Response model (Samejima, 1969), use different probability formulations that do not allow analysts to use threshold estimates to identify category disordering when it occurs (discussed further in Chapters 4 and 6).

Rasch Models for Rating Scale Analysis

In practice, researchers conduct rating scale analysis with three types of Rasch models: The Rating Scale Model (RSM), the Partial Credit Model (PCM), and RSM and PCM formulations of the Many-Facet Rasch Model (MFRM). Table 1.2 provided an overview of these models in terms of the types of data and rating scale analysis goals that each of them accommodates.

Before we discuss the use of these models to examine specific indicators of rating scale analysis, it is helpful to understand the basic characteristics of each model and how they can be used to provide an overview of the psychometric quality of survey responses. Accordingly, the remainder of this chapter includes a description of each model followed by a short example analysis with the CES-D scale data (see Chapter 1). Relatively more detail is provided for the RSM because this model shares many characteristics in common with the other models in this chapter. Statistical software scripts for the analyses are provided in the online supplement at **https://study.sagepub.com/researchmethods/qass/wind-exploring-rating-scale-functioning**. Chapter 3 provides a detailed demonstration of rating scale analyses using these models.

Rating Scale Model (RSM)

The RSM (Andrich, 1978), also known as the polytomous Rasch model, is a Rasch model for item responses in three or more ordered categories (e.g., $x = 0, 1, 2, \ldots, m$). As shown in Table 1.2, researchers use the RSM to analyze survey data when all of the items include the same set of response categories. The CES-D scale is an example of such an instrument because participants are presented with the same four rating scale categories for each item. In addition, researchers use the RSM to evaluate survey responses for evidence that they adhere to the fundamental measurement properties discussed in Chapter 1 and earlier in this chapter, including unidimensionality, local independence, and invariance. In the context of rating scale analysis, the RSM offers a relatively simple

procedure that researchers can use to evaluate rating scale functioning for an overall set of items (discussed further in Chapter 3).

The RSM is an extension of the dichotomous Rasch model for polytomous data. In log-odds form, the RSM states that the log of the odds for a response in category k, rather than in category $k - 1$ is determined by the difference between the participant location (θ), item location (δ), and rating scale category threshold locations (τ) on the logit scale that represents the latent variable:

$$ln\left(\frac{P_{ni(x=k)}}{P_{ni(x=k-1)}}\right) = \theta_n - (\delta_i + \tau_k). \tag{2.4}$$

As we saw with the dichotomous Rasch model, the RSM can also be expressed by converting the log-odds form to an exponent form, which describes the probability for a rating in a given rating scale category (category x) as:

$$P_{ni(x=k)} = \frac{\exp \sum_{k=0}^{x} [\theta_n - (\delta_i + \tau_k)]}{\sum_{j=0}^{m} \sum_{k=0}^{j} [\theta_n - (\delta_i + \tau_k)]}. \tag{2.5}$$

For Participant n on Item i where the maximum category is m, the probability for a rating in category x is expressed as the sum of the probabilities for the steps up to category x divided by the sum of the probabilities for all of the steps in the rating scale.

The RSM provides analysts with estimates of participant, item, and rating scale category threshold locations on the latent variable. We will consider the information that the RSM provides using an illustrative analysis with the CES-D scale data.

Application of the RSM to the CES-D Scale Data

The RSM was used to analyze participant responses to the CES-D scale, which includes 20 items with a four-category response scale recoded to $x = 0, 1, 2, 3$ (see Chapter 1). Applying the RSM to the CES-D scale goes beyond total-score-level analyses of the instrument to provide information about individual participant, item, and rating scale category locations on an interval-level scale that represents the construct. In addition, the RSM provides information about the quality of item responses from the perspective of invariant measurement

(Rasch, 1960). Most relevant to this book, the initial application of the RSM to the CES-D scale data provides information that can be used to evaluate rating scale functioning.

For the current illustration, the RSM was applied using the Facets software (Linacre, 2020), which uses Joint Maximum Likelihood Estimation (JMLE). Briefly, JMLE is an iterative procedure that involves calculating estimates for the model parameters (θ, δ, τ) using observed item responses. The procedure converts the observed probabilities in the data to measures on a log-odds scale, alternating between calculating item estimates and person estimates to find estimates that reflect the observed responses. For additional details on JMLE and other estimation procedures for Rasch models and IRT models, please see DeAyala (2009). Following typical Rasch estimation procedures, the mean of the item locations was set to zero logits to provide a frame of reference for interpreting the parameter estimates on the logit scale.

Preliminary Analysis: Model-Data Fit

Before interpreting the parameter estimates from the RSM in detail, it is important to examine the results for evidence that the data approximately reflect the expectations of the model. The purpose of this *model-data fit* analysis is to ensure that it is reasonable and appropriate to interpret the results before proceeding with further analysis, including rating scale analyses. This kind of analysis can be considered along the lines of checking assumptions for statistical models, such as checking the normality assumption in regression analysis. However, in the Rasch framework, model-data fit analysis is related to a theoretical framework that reflects requirements for measurement. Specifically, the Rasch approach begins with the hypothesis that item responses fit the Rasch model, and the data are fit to the model as an initial step. Then, residuals, or discrepancies, between model estimates and the data, are examined for evidence of substantial deviations from model requirements.

There are numerous techniques that researchers use in practice to evaluate adherence to Rasch model requirements, and it is beyond the scope of this book to explore Rasch model fit analysis in great detail. In this chapter, we consider model-data fit for Rasch models using three indices that are relatively straightforward to interpret. In practice, researchers often use all three of these indices to provide a comprehensive overview of model-data fit before they proceed with interpreting the model results.

Rasch model-data fit indicators are calculated using *residuals*, which are numeric summaries of the degree to which the observed responses for each person on each item (i.e., the actual survey responses) match the responses that we would expect to see for each person-item combination if the parameter estimates (person locations, item locations, and threshold locations) were accurate. Residuals are calculated for each person-item combination as follows:

$$Y_{ni} = X_{ni} - E_{ni}, \tag{2.6}$$

where X_{ni} is the observed response for person n on item i, and E_{ni} is the model-expected response for person n on item i. Model-expected responses are calculated using person location estimates (θ), item difficulty estimates (δ), and threshold locations (τ). Residuals are positive when the observed response was higher than the expected response (e.g., a person responded *Strongly Agree* when the expected response was *Agree*). Residuals are negative when the observed response was lower than the expected response (e.g., a person responded *Agree* when the expected response was *Strongly Agree*). Larger values of Y_{ni} indicate that there was a large difference between the observed response and the response that the model expected for a given person-item combination, and smaller values of Y_{ni} indicate a small difference.

Researchers can use residuals to explore many aspects of model-data fit. In this chapter, we use them to calculate three model-data fit indices that are relatively straightforward to interpret: (1) proportion of variance explained by model estimates; (2) correlations among item-specific residuals; and (3) numeric summaries of model residuals for items and persons. Graphical analyses that are also relevant to evaluating model-data fit for the RSM are included in Chapter 3.

Unidimensionality: Proportion of Variance Explained by Model Estimates

To begin, researchers often examine model results for evidence of adherence to the Rasch model requirement of unidimensionality (see Chapter 1). One unidimensionality evaluation procedure that is aligned with the Rasch framework is to evaluate how much of the variation in participant responses can be attributed to a single latent variable, such as depression in the context of the CES-D scale. To evaluate this property in practice, researchers can calculate the proportion of variance in responses that can be explained using Rasch model estimates.

This procedure is conducted automatically in the Facets software program (Linacre, 2020). It can also be approximated using three components; the supplemental materials demonstrate how to calculate these components in the R software packages that support Rasch model analyses. First, the variance of the original responses (V_O) is calculated using observed responses (X_{ni} in Equation 2.6). Then, the variance of *residuals* (V_R) is calculated using the Y_{ni} values from Equation 2.6. These values are combined to find the proportion of response variance attributable to Rasch model estimates: $(V_O - V_R)/V_O$.

For the simulated CES-D data, the approximate proportion of variance explained by Rasch model estimates was 24.55%. This value is greater than the minimum value of 20% that Reckase (1979) recommended for Rasch model analyses of potentially multidimensional scales—providing support for the use of the RSM to analyze the CES-D data.

Local Independence: Correlations Among Item-Specific Residuals

Next, one can evaluate the Rasch model requirement of local independence by examining correlations between the residuals that are associated with each item. The idea behind this analysis is this: If items are locally independent (thus satisfying the model requirement), there should be *no* meaningful relationships among the responses to individual items after controlling for the primary latent variable. Low absolute values of inter-item residual correlations (e.g., $|r| \leq 0.30$) provide evidence to support local independence (Yen, 1984). For the example CES-D scale data, the absolute values of the inter-item residual correlations were all less than or equal to $|r| = 0.04$—thus providing support for the use of the RSM to analyze these item responses. Practically speaking, this means that participants' responses to each item did not affect their responses to the other items in the scale after controlling for their level of depression.

Item- and Person-Specific Fit Analysis

Perhaps one of the most useful features of Rasch models, including the RSM, is item- and person-specific fit analysis. Item- and person-fit indices help analysts identify *individual items* and *individual persons* whose response patterns do not match what would be expected if the item response data adhered to the model requirements. Such analyses can be useful from a diagnostic perspective to improve data quality (e.g., to identify items that may be candidates for removal prior to further analysis), to identify individual participants whose responses

warrant additional exploration and consideration, to improve the quality of the instrument, to inform theory about the instrument or a sample, among other uses (see Chapter 1). In practice, many researchers use numeric summaries of item and person-specific residuals in the form of mean square error (MSE) statistics.

Specifically, one can examine unweighted or weighted means of standardized residuals for each item and person using outfit MSE statistics and infit MSE statistics, respectively (Smith, 2004). These statistics are calculated as follows. First, residuals are calculated for each item-person combination (Y_{ni}) using Equation 2.6. Then, standardized versions of the residuals (Z_{ni}) are calculated as:

$$Z_{ni} = Y_{ni} \Big/ \sqrt{W_{ni}}, \tag{2.7}$$

where W_{ni} is the variance of X_{ni}, calculated as:

$$W_{ni} = \sum_{k=0}^{M_i} (k - E_{ni})^2 p_{nik}, \tag{2.8}$$

where p_{nik} is the probability for a response in Category k from Person n on Item i, and M_i is the maximum category for Item i.

Outfit MSE statistics are unweighted means of standardized residuals specific to individual items or persons, calculated as:

$$\text{Outfit MSE} = \sum_{n=1}^{N} {z_{ni}}^2 \Big/ N. \tag{2.9}$$

Just as averages are sensitive to outliers in general statistical analyses, outfit MSE statistics are sensitive to extreme unexpected responses. In survey analyses, extreme unexpected responses occur when the observed response is much lower or higher than expected given model estimates. For example, an unexpected response could occur in the CES-D scale if a person with very mild or no depression responded *Most or All of the Time* on an item that would be considered a relatively strong indicator of depression. An unexpected response could also occur if a person with very severe depression responded *Rarely or None of the Time* to an item describing a common behavior among most of the participants, regardless of depression level.

Infit MSE statistics were developed to provide an indicator of model-data fit that is less sensitive to extreme residuals. These statistics

are calculated in a similar manner as outfit MSE, but they are weighted by response variance (W_{ni}):

$$\text{Infit MSE} = \sum_{n=1}^{N} W_{ni} Z_{ni}^{2} \Big/ \sum_{n=1}^{N} W_{ni}. \tag{2.10}$$

Because they are weighted, infit MSE statistics are less sensitive to extreme unexpected responses.

It is beyond the scope of the current text to discuss the interpretation of outfit and infit MSE in great detail; however, some basic guidance will be provided here. In contrast to some other statistics such as *t*-statistics, which have known distributions for specific sample sizes, there is no known sampling distribution for outfit and infit MSE statistics. As a result, they cannot be directly evaluated for statistical significance. Instead, many researchers use critical values (i.e., cut scores) based on practical guidance and empirical methods (e.g., bootstrap methods) to evaluate them in practical applications (DeAyala, 2009; Seol, 2016; Walker et al., 2018; Wolfe, 2013). In general, many researchers agree that values of outfit and infit MSE around 1.00 indicate acceptable fit to a Rasch model (Smith, 2004; Wu & Adams, 2013). Values that exceed 1.00 indicate more variation than expected in the responses associated with an item or person, and values that are less than 1.00 indicate less variation than expected. In many practical applications, researchers consider values of outfit and infit MSE that substantially exceed 1.00 as more cause for concern compared to low values of outfit and infit MSE (Linacre & Wright, 1994). When items or persons have notably high outfit and/or infit MSE statistics, analysts can examine the responses associated with the individual item or person in more detail for potential explanations. In some cases, it may be prudent to remove extreme misfitting items or persons from the data and reestimate model parameters in order to ensure meaningful interpretation of model results.

For the CES-D scale data, the mean outfit and infit MSE fit statistics were close to 1.00 for items (outfit MSE: $M = 0.99$, infit MSE: $M = 1.01$) and persons (outfit MSE = 0.99, infit MSE: $M = 1.00$). For items, the outfit MSE statistics ranged from 0.75 for Item 5 (*I had trouble keeping my mind on what I was doing*), which had responses that had the least amount of variation compared to model expectations, to 1.24 for Item 17 (*I had crying spells*), which had the most-frequent unexpected responses compared to model expectations. Infit MSE statistics ranged

from 0.73 for Item 5 to 1.34 for Item 17 (*I had crying spells*). Figure 2.4 illustrates the distribution of item fit statistics from the RSM.

Person-fit statistics are summarized visually in Figure 2.5. Person-infit MSE ranged from 0.36 for the participant with the most deterministic (i.e., predicable) response to 2.54 for the participant with the most frequent and substantial unexpected responses. Likewise, outfit MSE ranged from 0.37 to 1.98. Examination of the histograms of person-fit statistics suggests that the MSE fit statistics were around 1.00 for the majority of the sample. Although it is possible to explore item fit and person fit in more detail, these preliminary results are sufficient to proceed with further psychometric analyses with the RSM, including rating scale analysis.

Overall RSM Results

Figure 2.6 summarizes the results from the RSM analysis of the CES-D scale data using a Wright Map (i.e., a "variable map" or "item-person map"; see Wilson, 2011), which is a visual display that depicts the estimated locations for persons and items on a single linear scale that represents the construct. Wright maps are a key feature of Rasch measurement theory because they provide a concise summary of model results that capitalizes on the key strengths of Rasch models (Engelhard & Wang, 2020). Specifically, these displays illustrate the locations of individual items, persons, and rating scale categories on a single continuum that represents the construct (e.g., depression). As a result, they allow analysts to quickly visualize the location of individual elements within each facet (e.g., individual persons and items), the overall shape of the distributions of these elements, and to make comparisons between facet locations on the same scale. This visual summary of model estimates is invaluable for understanding and communicating the results from Rasch model analyses.

The first column in the Wright map (labeled "Logit") for the RS analysis of the CES-D scale shows the log-odds scale. This is the metric on which person, item, and rating scale category threshold locations were estimated. Low values indicate less-severe depression and high values indicate more-severe depression. The second column shows the distribution of person locations on the logit scale, where an asterisk symbol (*) represents 9 people and a period symbol (.) represents between 1 and 8 people. For persons, relatively low locations on the logit scale indicate persons with relatively mild depressive symptoms,

Figure 2.4 Histograms of Item Fit Statistics for the Rating Scale Model

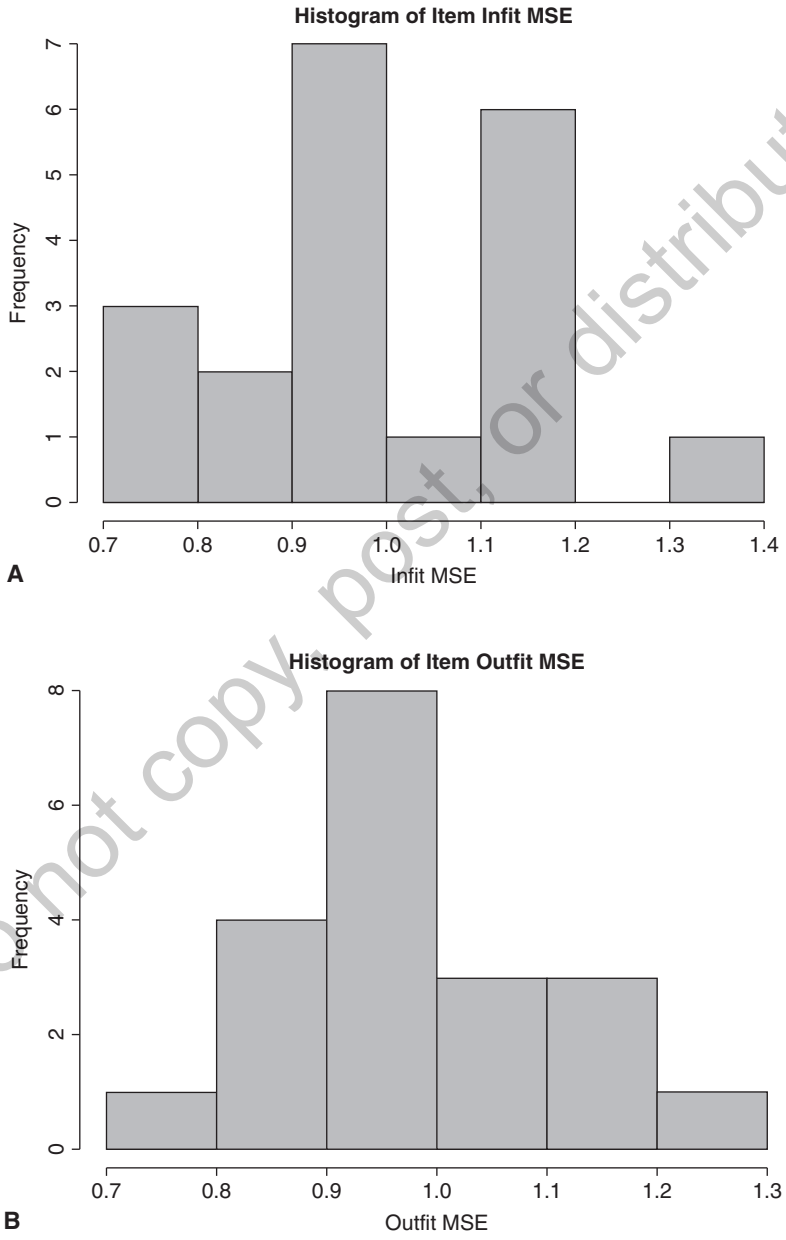


Figure 2.5 Histograms of Person-Fit Statistics for the Rating Scale Model

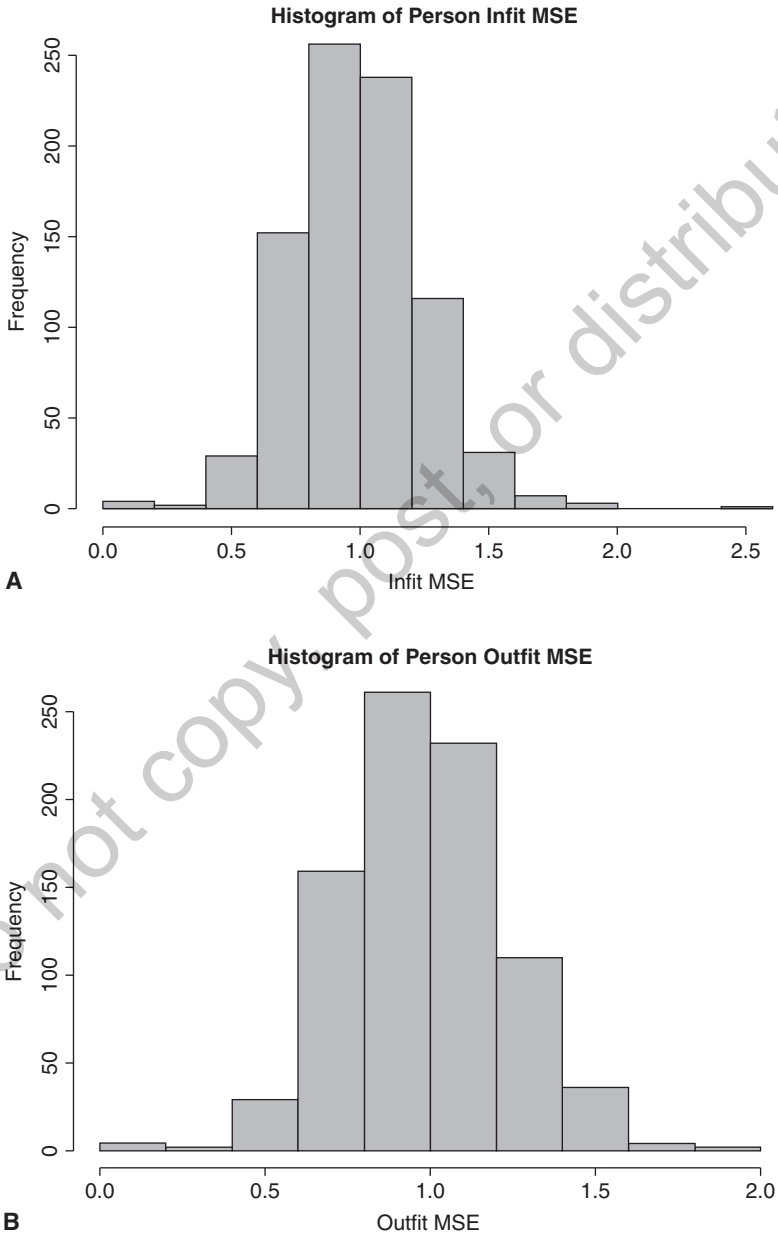
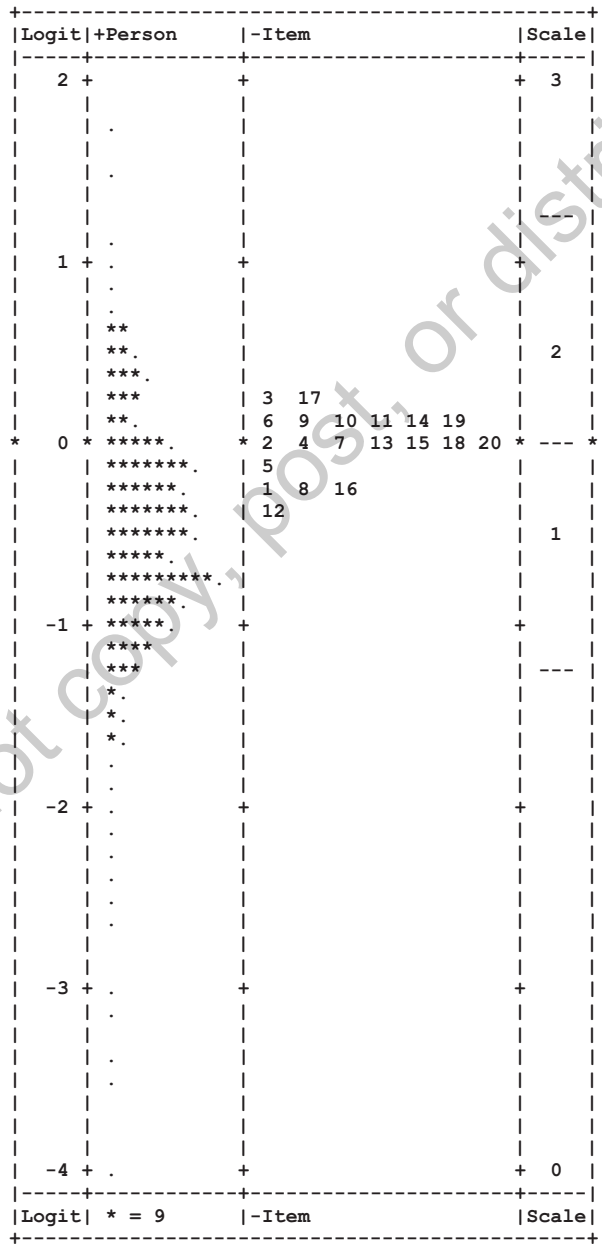


Figure 2.6 Wright Map for the CES-D Scale Data Based on the RSM
Note: In the Logit column, numbers indicate values on the log-odds scale. In the Person column, asterisks (*) indicate 9 people and periods (.) indicate between 1 and 8 people. In the Item column, numbers indicate item numbers from the CES-D scale. In the Scale column, numbers indicate category numbers from the recoded CES-D scale, which ranges from 0 to 3. Dashed horizontal lines indicate thresholds between adjacent categories in the scale



and relatively high locations on the logit scale indicate persons with relatively severe depressive symptoms. These results indicate that the average person location was equal to −0.54 logits, and that the distribution of person locations was approximately normal or bell-shaped.

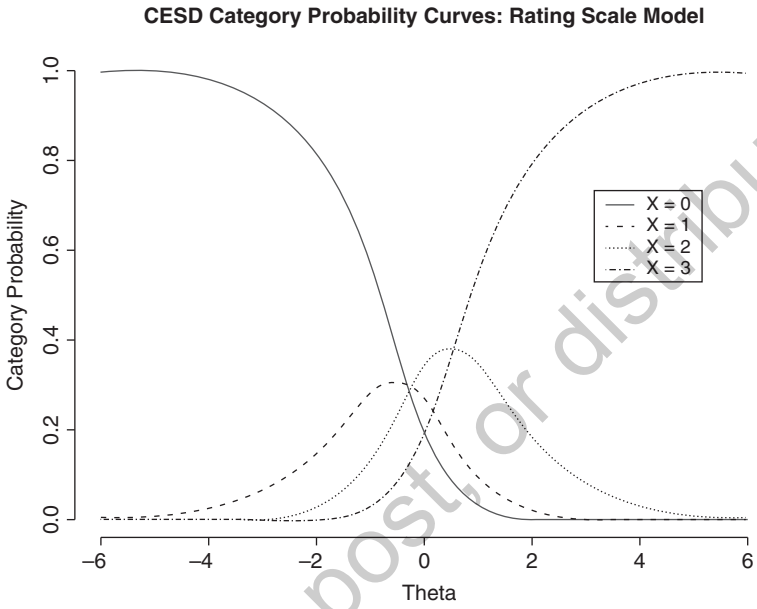
The third column of Figure 2.6 shows the locations of the overall item estimates (δ) based on the RSM, with item numbers used to mark estimated item locations. For items, low locations on the logit scale indicate that an item requires participants to experience relatively mild depressive symptoms to report that they frequently experience or perform the stated behavior, and high locations on the logit scale indicate that an item requires participants to experience relatively severe depressive symptoms to report that they frequently experience or perform the stated behavior. In the estimation procedure, the average item location was set to zero logits to provide a frame of reference for interpreting the logit scale. The results from this analysis indicate that on average, the persons had lower locations on the logit scale relative to items—indicating that the participants exhibited relatively low levels of depression.

The final column of Figure 2.6 shows the estimated locations for the rating scale category thresholds using dashed horizontal lines between numeric labels for the category numbers ($x = 0, 1, 2, 3$). Because the CES-D rating scale includes four categories, there are three rating scale category threshold estimates (τ_1, τ_2, τ_3). The estimated threshold locations in logits were as follows: $\tau_1 = -0.30$, $\tau_2 = -0.25$, and $\tau_3 = 0.55$. The distance between the first and second thresholds (τ_1 and τ_2) is very small (approximately 0.05 logits). To further illustrate these results, Figure 2.7 shows rating scale category probability curves for the CES-D rating scale, as estimated with the RSM. Both the numeric and graphical results indicate that the second rating scale category does not have a distinct range on the logit scale at which it is the most probable. We consider these results in more detail in Chapter 3.

Partial Credit Model (PCM)

The PCM (Masters, 1982) is similar in many ways to the RSM: It is a polytomous Rasch model for item responses in three or more ordered categories (e.g., $x = 0, 1, 2, \ldots, m$) that provides researchers with estimates of person locations (θ), item locations (δ), and rating scale category threshold locations (τ) on a linear scale that represents a latent variable. The major difference from the RSM is that in the PCM,

Figure 2.7 Rating Scale Category Probability Curves for the CES-D Scale Data Based on the RSM



threshold locations are calculated separately for each item. As shown in Table 1.2, this feature allows researchers to use the PCM when items have different response scales, and when they want to evaluate rating scale functioning separately for each item. We discuss choosing between the RSM and PCM in more detail in Chapters 3 and 6.

In log-odds form, the PCM states that the log of the odds that Participant n gives a rating in Category k rather than in category $k-1$ is determined by the difference between the Participant's location on the construct (θ) and the combination of the item location parameter (δ_i) with the threshold parameter (τ_k), specific to Item i:

$$\ln\left(\frac{P_{ni(x=k)}}{P_{ni(x=k-1)}}\right) = \theta_n - \delta_{ik} \tag{2.11}$$

In exponent form, the PCM expresses the probability for a rating in a given rating scale category (category x) is stated as:

$$P_{ni(x=k)} = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_{ik})}{\sum_{j=0}^{m} \sum_{k=0}^{j} (\theta_n - \delta_{ik})}. \quad (2.12)$$

In the PCM, the item location parameter (δ_i) is combined with the threshold parameter (τ_k) for each item such that δ_{ik} is the location on the logit scale at which there is an equal probability for a rating in category k and category $k - 1$, specific to item i. This means that each item has its own unique set of threshold estimates, as illustrated in Figure 1.3.

When the PCM is applied to item response data, it provides estimates of participant, item, and rating scale category threshold locations specific to each item. These results can be used to examine a variety of psychometric properties, including rating scale functioning. The PCM is particularly useful for rating scale analysis because it facilitates an examination of rating scale functioning specific to each item in a scale. This information can be useful for identifying individual items for which rating scales are not functioning as expected. Chapter 3 provides a more in-depth exploration of the use of the PCM for this purpose.

Application of the PCM to the CES-D Data

The PCM was used to analyze participant responses to the CES-D scale using the Facets software with item locations centered at zero logits. As noted in the demonstration of the RSM, it is important to evaluate the degree to which item responses approximate Rasch model requirements before interpreting parameter estimates in detail. The model-data fit analysis results for the PCM were similar to those reported earlier for the RSM. Specifically, the PCM estimates explained 25.2% of the variance in observed responses and all of the inter-item residual correlations had absolute values equal to or less than or equal to $|r| = 0.03$. For items, the largest values of outfit MSE and infit MSE were observed for the recoded version of Item 12 (*I was happy*; outfit MSE = 1.26; infit MSE = 1.22) and the recoded version of Item 16 (*I enjoyed life*; outfit MSE = 1.24; infit MSE = 1.19). The lowest values of the MSE statistics were observed for Item 5 (*I had trouble keeping my mind on what I was doing*; outfit MSE = 0.88; infit MSE = 0.89).

Overall, the results from the PCM analysis of the CES-D scale data are similar to those from the RSM. Figure 2.8 summarizes the results

using a Wright Map in the same format as presented in Figure 2.6 for the RSM. The major difference is that the rating scale category thresholds are estimated separately for each item; these item-specific thresholds are shown in the 20 columns to the right of the overall item locations. The thresholds that correspond to each item are presented in a column labeled "S" for *scale* followed by the item number. Examination of the threshold estimates indicates that there are differences in the structure of the rating scale across the CES-D items. We explore these differences in detail using rating scale category probability curves and other indicators in Chapter 3.

Extending the Rating Scale and Partial Credit Models: The Many-Facet Rasch Model (MFRM)

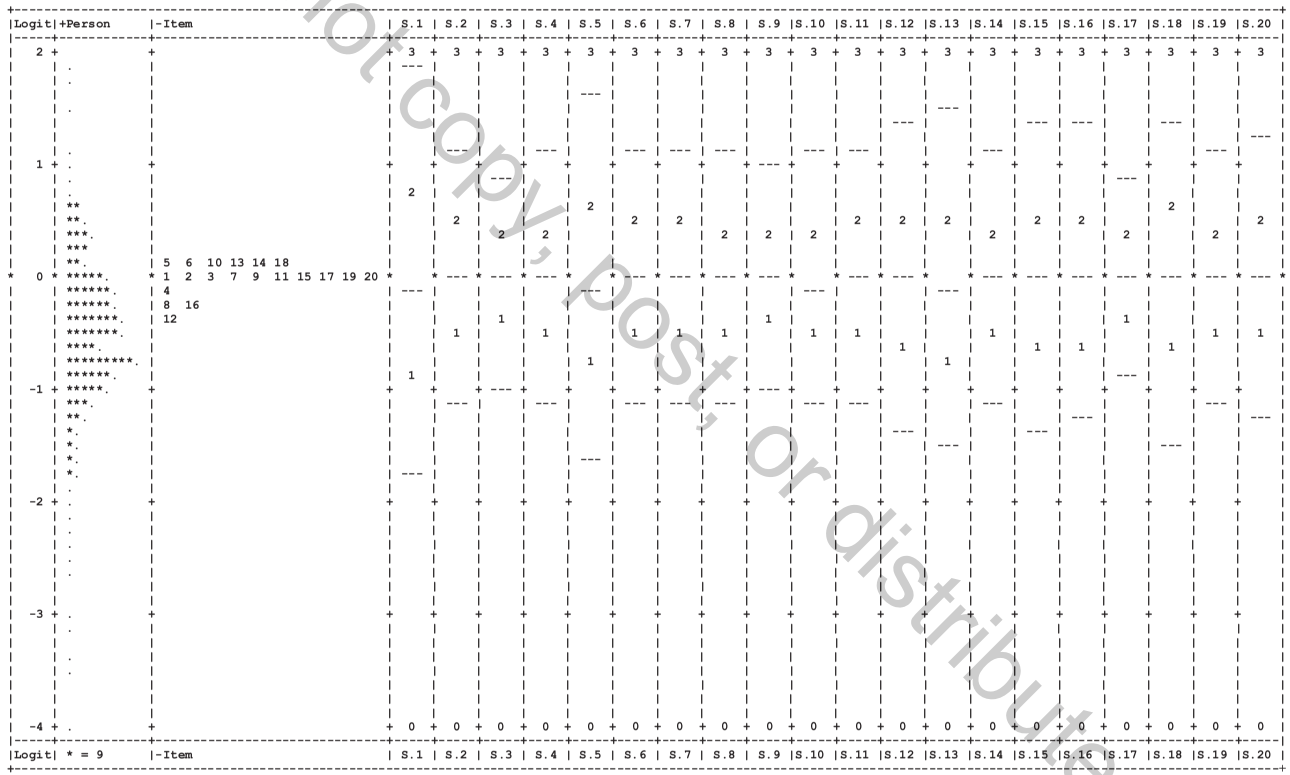
In many assessment contexts, additional components of the assessment system besides persons and items contribute to item responses in important ways. Linacre (1989) proposed the Many-Facet Rasch Model (MFRM) as a flexible extension of Rasch models that allows researchers to include explanatory variables ("facets"). The MFRM is similar to a logistic regression model where researchers can examine the relationship between an independent and dependent variable controlling for other variables.

The MFRM was originally proposed in the context of performance assessments in which raters (i.e., judges) score participant performances. In this context, raters can be included as a facet to estimate rater severity levels in the same frame of reference as participants and items. This model also allows researchers to examine the impact of differences in rater severity on the estimates of student achievement, item difficulty, and rating scale category thresholds. Beyond rater-mediated assessments, the MFRM can be applied to a variety of contexts in which it is useful to include additional explanatory facets besides item and person locations. For example, many researchers use MFRMs to estimate logit-scale locations related to item or person features, such as demographic subgroups of persons, item types, or administrations of an assessment procedure in longitudinal designs (e.g., Gordon et al., 2021; Ho, 2019; Primi et al., 2019; Toffoli et al., 2016).

The MFRM is a flexible model that can be used to extend each of the Rasch models that have been discussed so far in this book (dichotomous, RSM, and PCM). The MFRM can also be used to extend Rasch models that are not described in this book, including the binomial trials

Figure 2.8 Wright Map for the CES-D Scale Data Based on the PCM

Note: See the note for Figure 2.6 for descriptions of the Logit, Person, and Item columns. Separate scale columns are presented for each item, labeled "S." followed by the item number from the CES-D scale



and the Poisson counts Rasch models (see Wright & Mok, 2004). Given its flexible nature, there is no single formulation of the MFRM that appears in the literature. Instead, researchers specify their own unique set of facets and add them to an appropriate Rasch model to reflect their measurement context. As shown in Table 1.2, this feature allows researchers to use the MFRM to examine rating scale functioning related to explanatory variables that are unique to each assessment context. We discuss this use of the MFRM in Chapter 3.

In the context of rating scale analysis, the MFRM can be used to extend the RSM and the PCM. A general form of a Rating Scale model formulation of the MFRM (RS-MFRM) can be stated in log-odds form as:

$$\ln\left(\frac{P_{n(x=k)}}{P_{n(x=k-1)}}\right) = \theta_n - \sum_{\text{facets}} \varepsilon - \tau_k, \tag{2.13}$$

where θ_n and τ_k are defined as in the RSM and $\sum_{\text{facets}} \varepsilon$ is a linear combination of the researcher-specified facets that reflect aspects of the assessment system. The estimate of the person's location on the latent variable (θ_n) is controlled (i.e., adjusted) for the facets included in $\sum_{\text{facets}} \varepsilon$. For example, a researcher might specify a RS-MFRM that includes facets for participants, items, and participant education-level subgroups. This allows the analyst to examine the probability for a response controlling for differences related to participant education level. Stated in log-odds form, this RS-MFRM is:

$$\ln\left(\frac{P_{nji(x=k)}}{P_{nji(x=k-1)}}\right) = \theta_n - \gamma_j - \delta_i - \tau_k, \tag{2.14}$$

where θ_n, δ_i, and τ_k are defined as in the RSM, and γ_j is the logit-scale location for participant subgroup (e.g., education level) j. Researchers may use the RS-MFRM when they want to examine rating scale functioning for an overall set of items while controlling for an explanatory facet such as education level (see Table 1.2).

Similarly, a PCM formulation of the MFRM (PC-MFRM) with the same facets could be specified as:

$$\ln\left(\frac{P_{nji(x=k)}}{P_{nji(x=k-1)}}\right) = \theta_n - \gamma_j - \delta_i - \tau_{ik}, \tag{2.15}$$

In the PC-MFRM, the threshold parameter includes subscripts for items and rating scale categories (τ_{ik})—indicating that separate rating

scale category thresholds are estimated for each item in the same manner as was presented for the PCM. This would allow researchers to examine rating scale functioning for individual items while controlling for an explanatory facet such as education level.

The PC-MFRM can also be specified to allow researchers to examine rating scale functioning specific to the levels of an explanatory facet. For example, researchers may wish to evaluate the degree to which a rating scale functions in a comparable way between participants with different levels of education. This type of model could be specified by changing the subscript on the threshold parameter so that thresholds vary across the j education level subgroups:

$$\ln\left(\frac{P_{nji(x=k)}}{P_{nji(x=k-1)}}\right) = \theta_n - \gamma_j - \delta_i - \tau_{jk}, \tag{2.16}$$

In this specification, the threshold parameter includes subscripts for participant subgroups and rating scale categories (τ_{jk})—indicating that separate rating scale category thresholds are estimated for each subgroup. In practice, the PC-MFRM is particularly useful for rating scale analysis because it facilitates an examination of rating scale functioning specific to each level of a facet of interest in an assessment system (see Table 1.2). This information can be useful for identifying individual levels of facets for which rating scales are not functioning as expected, and to examine the consistency of rating scale functioning across levels of facets (e.g., across participant subgroups). Chapter 3 provides a more in-depth exploration of the use of the PC-MFRM for this purpose.

Application of the PC-MFRM to the CES-D Data

Next, we will apply the PC-MFRM given in Equation 2.16 to analyze participant responses to the CES-D scale. In the CES-D scale data examined in this book, participants' education level was reported using six categories: (1) eighth grade or less, (2) some high school, (3) high school or high-school graduate equivalent, (4) completed some college or two-year degree, (5) completed four-year degree, and (6) graduate or professional degree. As in the previous analyses presented in this chapter, the mean of the item locations was set to zero logits.

In addition, the logit scale location for the eighth grade or less education level subgroup was fixed to zero logits, and the remaining education level subgroup locations were estimated freely. This provided

a frame of reference for interpreting and comparing participant subgroups on the logit scale.

As noted earlier in the demonstration of the RSM and the PCM, it is important to evaluate the degree to which item responses approximate Rasch model requirements before interpreting parameter estimates in detail. For the PC-MFRM, the fit analysis results generally agreed with those from the RSM and PCM analyses. Specifically, PC-MFRM estimates explained 24.55% of the variance in observed responses, and the absolute value of each of the inter-item residual correlations was less than or equal to $|r| = 0.04$.

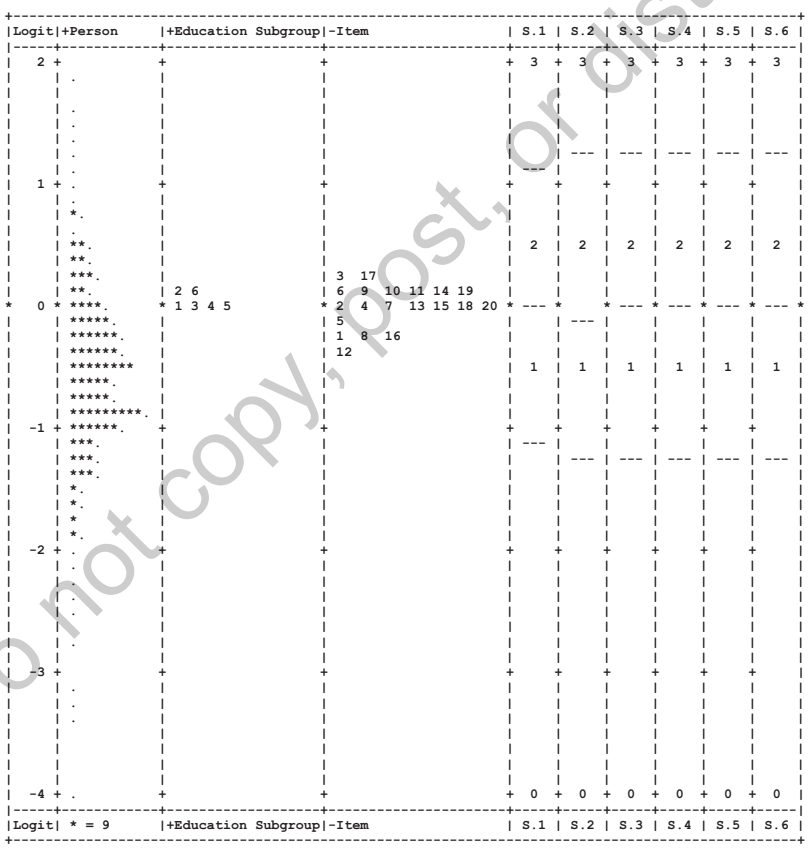
Figure 2.9 summarizes the results using a Wright Map in the same format as presented earlier for the RSM and the PCM, with an additional column (the third column) that shows the logit scale locations for each education-level subgroup. In addition, rating scale category thresholds are estimated separately for each subgroup; these subgroup-specific thresholds are shown in the six columns to the right of the overall item locations. The thresholds that correspond to each item are presented in a column labeled "S" for *scale* followed by the subgroup number. Examination of the threshold estimates indicates that there are differences in the structure of the rating scale across the CES-D items. We explore these differences using rating scale category probability curves and other indicators specific to each education subgroup in detail in Chapter 3.

Chapter Summary

This chapter began with a brief overview of Rasch measurement theory as a framework characterized by clear requirements that reflect measurement properties in the physical sciences. Then, two popular Rasch models for rating scale analysis were introduced with example applications using the CES-D data: The Rating Scale Model (RSM; Andrich, 1978) and the Partial Credit Model (PCM; Masters, 1982). Both of these models provide estimates of person locations, item locations, and rating scale threshold locations on a linear scale that represents a latent variable. The major difference between the models is that the PCM specifies rating scale thresholds separately for each item. Researchers may choose the RSM when they want an overall summary of rating scale functioning without item-specific details. Researchers may choose the PCM when they want item-specific details about rating scale functioning (see Table 1.2, discussed in more detail in Chapters 3

Figure 2.9 Wright Map for the CES-D Scale Data Based on the PC-MFRM

Note: See the note for Figure 2.6 for descriptions of the Logit, Person, and Item columns. In the Education Subgroup column, numbers indicate subgroups of participants with different levels of education: (1) eighth grade or less, (2) some high school, (3) high school or high school graduate equivalent, (4) completed some college or two-year degree, (5) completed four-year degree, and (6) graduate or professional degree. Separate scale columns are presented for each Education Subgroup, labeled "S." followed by the subgroup number as indicated in the Education Subgroup column



and 6). Next, the Many-Facet Rasch Model (MFRM; Linacre, 1989) was presented and illustrated with the CES-D data as an extension of Rasch models that can be customized to reflect a variety of contexts and data analysis purposes that are relevant for rating scale analysis. Building on this content, Chapter 3 presents and illustrates techniques for exploring rating scale functioning using polytomous Rasch models.