Chapter 3
Applying the Rasch Rating Scale Method to Questionnaire Data

Christine DiStefano and Ning Jiang

Abstract This chapter provides an introduction to the Rasch rating scale model (RSM) and provides a primer of how to use the methodology when analyzing questionnaires. The work includes a discussion of best practices for using the RSM, how to evaluate item and person fit, and how to use the information to build a psychometrically sound scale. An applied example is provided to assist researchers with their decision making.

Keywords Rasch rating scale · Wright map · Latent construct · Probability · Infit · Outfit

Introduction

In the social sciences, questionnaires are frequently used to collect data about a variety of educational, social and behavioral construct in which responses are thought to reflect evaluations about an area of interest. Use of survey instruments in general afford many advantages to the research community including ease of distribution options through various modalities (e.g., telephone, mail, paper-pencil, on-line); the opportunity to collect a wide variety of information, from demographic characteristics to sensitive issues; and the ability to collect self-report data or proxy data (i.e., where persons complete information and reflections about someone other than themselves) from respondents. Many scales are available to use on questionnaires including items which as respondents to provide rankings on a checklist of stimuli, forced choice options, and even open-ended questions. The most popular types of survey items typically include closed-ended scales such as Likert scaled items or performance rating scales.

Ordinal scales allow respondents to select a rating according along a continuum. These scales have many advantages, such as producing data which are relatively easy to collect, summarize, and report (Fink, 2012). Likert scales are by far the most used method for collecting data, as the scales are easily adaptable to many situations,

C. DiStefano (✉) · N. Jiang
University of South Carolina, Columbia, SC, USA
e-mail: DISTEFAN@mailbox.sc.edu

© Springer Nature Singapore Pte Ltd. 2020

M. S. Khine (ed.), *Rasch Measurement*,
https://doi.org/10.1007/978-981-15-1800-3_3

with choices of anchors that allow researchers to collect data on a wide variety of perspectives such as frequency, intensity, agreement, and likelihood (Fowler, 2013). Further, the number of scale points may be adjusted to include a greater number of scale points (producing more continuous-like data), adding a middle or neutral response category, and using few categories or even pictures to collect data from children (Fink, 2012; Fowler, 2013; Nardi, 2018).

Often, researchers use responses from an ordinal scale to represent a construct of interest by summing item responses to create a total scale score. This assumes that the items have at least interval level properties, that is, that the distance between categories is the same for all respondents. In addition, the same (unit) weight is given to all items (DiStefano, Zhu, & Mindrila, 2009). However, summing responses assumes at least interval level of data—and this assumption may be questionable when ordinal data are present (Bond & Fox, 2007; Iramaneerat, Smith, & Smith, 2008). Further, summed scores do not give additional consideration to items that may vary due to the item's placement relative to the construct (i.e., difficulty value). Finally, characteristics of items are not typically examined beyond descriptive information, such as the number of respondents per category.

As a better alternative, there are applications within the Rasch family that can be used to examine ordinal data (Smith, Wakely, De Kruif, & Swartz, 2003). The Rasch Rating Scale Model (RSM) is an optimal method for examining providing information about data fit to the model, information about characteristics of items and samples such as dimensionality of the measure, use of the rating scale, and coverage of the latent dimension (e.g., Kahler, Strong, & Read, 2005; Thomas, 2011). The purpose of this chapter is to introduce researchers to characteristics of the RSM including: the structure of the model, assumptions needed for accurate assessment, and how to evaluate results from RSM analyses. We provide information concerning these objectives and present an applied example to illustrate these characteristics in practice. The chapter closes by including additional applications for using the RSM for scale development, predicting latent scores, as well as suggestions for future research in this area.

Rasch RSM Methodology Overview

In general, Rasch methods refer to a family of mathematical models that compute the probability an individual will respond favorably to an item given the item's characteristics. The Rating Scale Model (RSM) is a specialized Rasch model for polytomously scored items; however, it follows the same perspectives (i.e., common metric, sample free measurement, linear latent scores) as with Rasch with dichotomous data (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). In addition, the goal of RSM is the same as other Rasch models—to provide scores for each person and each item on a common, interval-level (i.e., logit or log-odds) scale.

The Rasch RSM is a specialized model for use with ordinal data, such as responses from a Likert scale. The model incorporates a threshold value into the item estimation

process. For polytomous data, the number of thresholds is equal to the number of scale categories (k) minus 1. For example, a four-point scale would have three thresholds—or three points which cut the distribution of responses into four ordered categories (Bond & Fox, 2007). The threshold can be thought of as the point which moves a rating from one category into an adjacent category on the Likert scale. Thus, the threshold τ_{ki} partitions the continuum into set "categories" above and below its location. The threshold value corresponds with the location on a latent continuum at which it is equally likely a person will be classified into adjacent categories, and, therefore, likely to obtain one of two successive scores. Considering an item (i) with four categories, the first threshold of the item, τ_{1i} is the location on the continuum at which a respondent is equally likely to obtain a score of 0 or 1, the second threshold is the location at which a respondent is equally likely to obtain a score of 1 and 2, etc., through the k categories included with the ordered scale (Smith et al., 2003).

The RSM formula can be summarized as:

$$\Pr\{X_{ni} = x\} = \frac{\exp \sum_{k-0}^{x} (\beta_n - (\delta_i - \tau_k))}{\sum_{j=0}^{m} \exp \sum_{k=0}^{j} (\beta_n - (\delta_i - \tau_k))},$$

where β_n is the level of the construct for a given person, δ_i is the difficulty of item i and τ_i is the kth threshold location of the rating scale which is the same to all the items, m is the maximum score. The resulting quotient is a probability value showing the likelihood that a category will be selected given both the difficulty of the item and the individual's level of the construct under study. These probabilities can be transformed into a logit score by taking the natural odds log value. The logit score will vary if the probability is computed across all respondents for an item (item logit) or across items to compute the score for an individual (person logit).

Assumptions. The Rasch RSM includes the same assumptions as with the dichtomous Rasch model that should be met for accurate parameter estimation. These assumptions include: (1) construct unidimensionality, (2) a monotonic scale (i.e., higher latent scores represent a higher level of the latent construct), and (3) that the items fit the Rasch model (Bond & Fox, 2007; Sick, 2010). These three assumptions can be tested in the same manner with RSM as with dichtomous Rasch models. For example, unidimensionality with RSM is assessed using an unrotated Principal Component Analysis of standardized residuals to determine if there is additional variance to be explained after the latent construct has been extracted (Bond & Fox, 2007). Additional requirements (described below) are needed when using the RSM. If the requirements underlying RSM are met, the model offers the same benefits as with other Rasch models: (1) a common interval level metric for calibrated item and person measures, (2) fit statistics to evaluate items and persons which do not align with the Rasch model (i.e., misfit), (3) estimation of projected ratings for the latent construct, and (4) evaluation of the breadth of item coverage of the latent construct.

Rating Scale Diagnostics

A major benefit of the Rasch RSM is the ability to examine characteristics of category performance, frequency of category use, and interpretation of the scale (Bond & Fox, 2007). These investigations should be conducted at the start of a Rasch RSM to ensure that the scale and the categories are functioning properly. If the scales are not functioning as expected, the result is uninterpretable data. Therefore, the first step for the applied researcher utilizing RSM is to investigate rating scale performance, and, if necessary, to make improvements to the scale. The primary objective is to obtain a rating scale that produces the highest quality data for measuring the construct of interest.

Category Usage. The first step in RSM is to examine how respondents are using the categories of the rating scale. This analysis is largely descriptive and examines both the category frequencies and average measures per category. The category frequency provides the distribution of responses, indicating the number of respondents selecting a given category, summed for each category across all items on the questionnaire.

As noted by Bond and Fox (2007), researchers should investigate the shape of the distribution as well as the number of respondents per category. The shape of the distribution (e.g., normal, bimodal, uniform, skewed) provides information about the construct under study. In the social sciences, non-normal distributions are likely to be the standard rather than the exception (Finney & DiStefano, 2013; Micceri, 1989). While slight distributional anomalies are likely to be present, estimation problems may arise if the distribution is irregular, such as highly skewed or kurtotic.

In addition, the observed count in each category provides evidence of the category usage of respondents. Categories with low numbers of respondents do not provide sufficient information to allow stable estimation. Further, categories with few responses illustrate unneeded or even redundant categories, and may be collapsed into adjacent categories. It is recommended that each response category (k) has a minimum frequency of 10 respondents (Smith et al., 2003).

Another characteristic which is evaluated is the average measure value associated with each threshold. The average measure is the average of the ability estimates all persons who chose that particular response category with the average calculated across all observations in a given category. This value can be used to examine if the scale is performing adequately, including an increasing scale (e.g., persons with higher levels of the latent construct are expected to endorse higher levels of the scale).

Along with the average measure values, average Outfit measures associated with each category may also be examined using "standard" fit criteria (i.e., values less than 2.0). This investigation provides information about the quality of the rating scale. Outfit measures which are greater than 2.0 show that there is typically more misinformation than information, meaning that the category is introducing noise into the analyses (Bond & Fox, 2007; Linacre, 2004).

Threshold Values and Category Fit. Category performance may be evaluated by investigating the threshold values (or step calibrations) to determine if respondents are using the categories as expected. It is expected that rating scale categories increase in difficulty of endorsement, and that the thresholds for each item are ordered (Iramaneerat et al., 2008; Smith et al., 2003). The step measure parameter defines the location between categories, which should increase monotonically with categories. Disordering of step measures occurs when the rating scale does not function properly (Linacre, 2002). Thresholds should increase by at least 1.4 logits between categories but not more than 5 logits to avoid large gaps (Linacre, 1999).

A probability curve can be used to examine if the is performing optimally through visual inspection. This is a curve illustrating the probability of responding to a particular category given the difference in estimates between the person's level of the construct and the difficulty of the item ($\beta - \delta$). The curve plots the probability of responses on the y-axis and the person measure scores on the x-axis; individual curves for each category are presented in the body of the figure. When examining curves, researchers should note the shape and height of a given curve. Curves that are "flat" cover a large portion of the construct; however, these curves may also illustrate redundant or unneeded categories. Each curve should show a "peak". This suggests the category is the most probable response category for at least some portion of the construct measured by the questionnaire (Bond & Fox, 2007).

Figure 3.1 provides an illustration of a probability curve. Here, it can be seen that there is a four-category scale, with three threshold values noted by the asterisk (*)

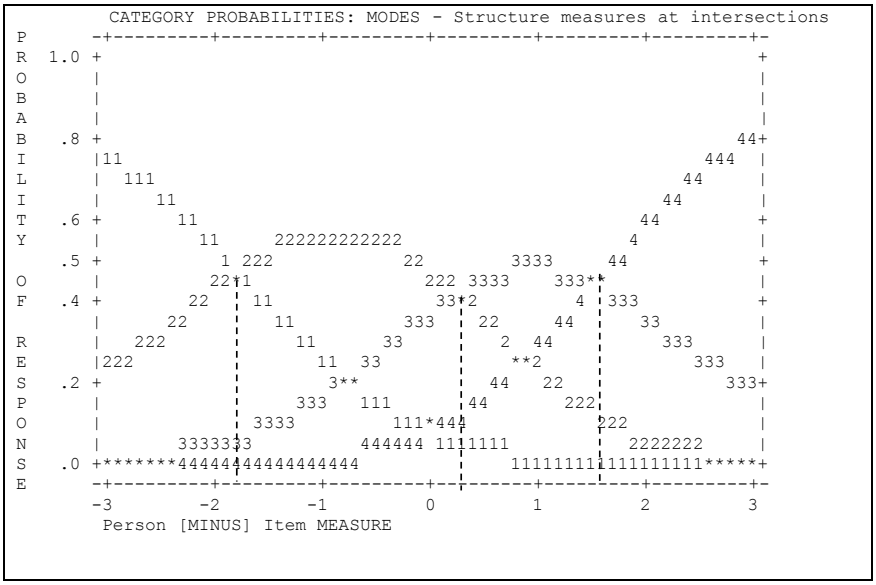


Fig. 3.1 Intensity of physical activity participation scale, 4 categories (from DiStefano et al., 2016). *Note* Threshold values are denoted by dotted lines

values. As noted below, each category displays a maximum peak, showing that it is the optimal response category (on average) for some respondents along the continuum. In addition, the dotted line shows the average construct score for a given threshold value. For example, the (approximate) threshold value between categories 1 and 2 is roughly −1.8. This can be interpreted as respondents with a person measure score that is lower than −1.8 would likely select category 1; persons with scores between −1.8 and (approximately) 0.3 would be expected to select the 2nd category. In this way, the expected category which a respondent would select, based on their overall measure, can be evaluated using the probability curve.

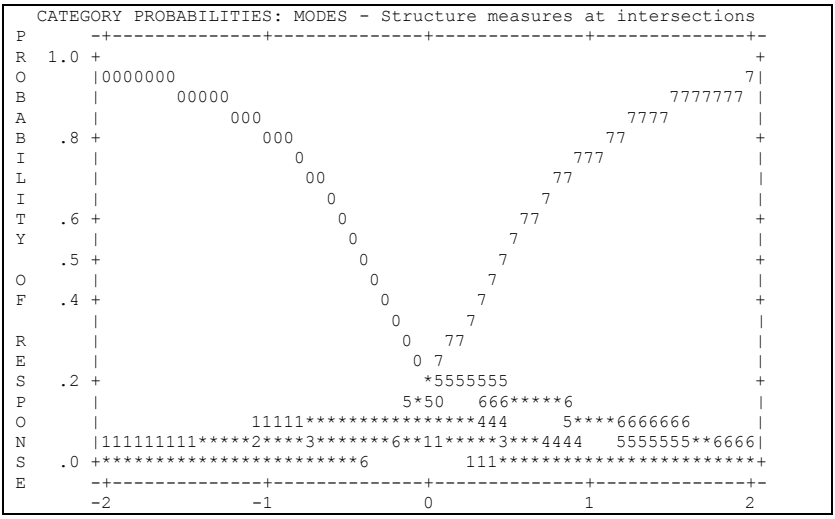
Using RSM Information for Scale Revision. Scale categories which are not utilized or well understood by respondents—such as: scales which include a mix of negatively and positively worded items, unclear wording on a questionnaire, or including too many response categories may show aberrant patterns. For example, Fig. 3.2a shows a scale which was originally conceptualized as an eight-category scale; however, as seen below, many of the categories were not sufficiently used, resulting in lower than recommended frequencies per category and disordered step values.

Here, the scale should be recoded to eliminate misfit and to ensure that the assumptions needed for RSM estimation are obtained. For scale development situations, this investigation can also suggest revisions to the ordinal scale to be used with future administrations of the questionnaire. Figure 3.2b recodes the same scale with three ordered categories, collapsing the scale from the original 0–7 to recoded values of 0 (0–1 from the original scale), 1 (2–3), and 2 (4–7). As can be seen here, recoding the eight-category measure to a three-category scales eliminates problems, producing a scale which functioned acceptably (i.e., no misfit). This can be observed by noting the ordered threshold values (*) between categories and a definite peak for each category included on the scale. As a reminder, any scale revisions should be conducted during the questionnaire's piloting stage to ensure that the best measurement can be obtained.

Visual Representations of the Latent Dimension

Coverage of the latent dimension and expected responses may be examined using Wright Maps and Expected Probability Maps (Bond & Fox, 2007). These maps are similar to the ones presented with other Rasch analyses, however, the plots may be helpful to interpret when conducting RSM. First, a Wright map (or Person-Item map) may be examined to determine the concordance between estimated ability levels of a sample of examinees relative to item difficulty values. These maps typically provide a picture of both calibrated abilities and difficulties along a continuum. For person and item distribution of scores, the mean (M) is provided in the center of the distribution with one (S) and two (T) standard deviations from the mean noted. Person-item maps are very useful in questionnaire development for many reasons such as identifying item redundancy and ensuring that the items on the questionnaire are focused at the

(a)



(b)

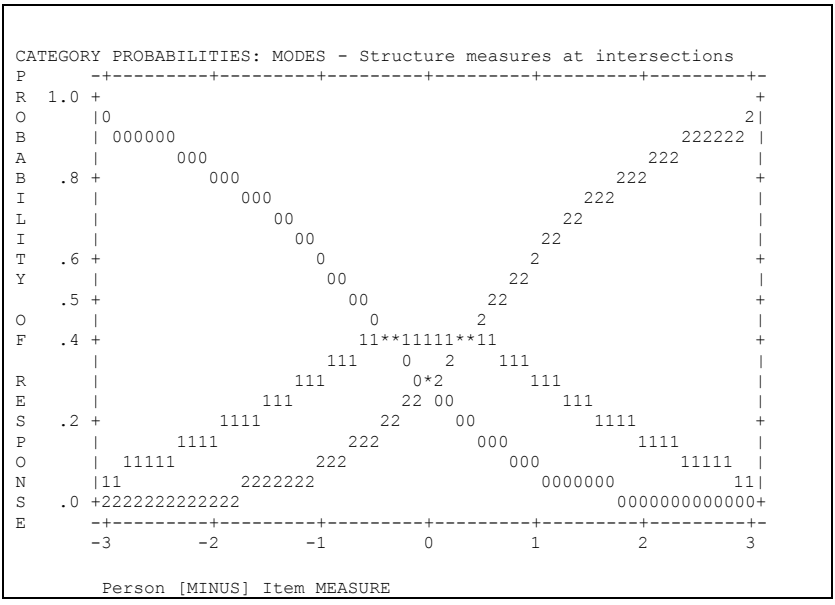


Fig. 3.2 **a** Non-optimal performing Likert scale, eight ordered categories. **b** Ordinal RSM scale recoded to three ordered categories

target level. While the same Wright map is typical in Rasch analyses, with RSM, there are "multiple" threshold estimates that are produced. Thus, multiple threshold levels for a given item are provided. The graphs allow for an examination of category endorsement relative to the distribution of levels of the construct under study.

Another useful graph for RSM analyses is the Expected Response Probability graph. This graph illustrates the expected responses that would be selected for each item on the ordinal scale, given different levels of the latent variable under study. The graph provides a continuum of person measures along the x-axis; along the y-axis are questionnaire items, ordered according to item difficulty values. The rating scale values (e.g., 0, 1, and 2) are provided with colons (threshold values) noted. The colons show where a respondent would mark the next highest category on the rating scale if the threshold is surpassed, given the person level of the latent variable. Expected Response Probability graphs may be useful to examine how expected responses to determine how examinees at targeted levels may respond to the rating scale and also of interest for test users to examine to identify what expected responses to scale items may be for different ability levels of respondents.

Illustrative Example

To assist researchers with interpretation of the decisions involved with a Rasch RSM, we provide an example to highlight information and choices that may be encountered when analyzing questionnaire data. The example utilizes the Externalizing Problems scale from the Pediatric Symptoms Checklist, 17-item screener (PSC-17, Gardner et al., 1999). The PSC-17 is a short version of the full PSC measure (35-items) which is often used to measure children's emotional and behavioral risk (Jellinek et al., 1988). The screener consists of 17 items, measuring three kinds of mental health problems: internalizing problems, attention problems, and externalizing problems. Both the Internalizing Problems subscale and the Attention Problems subscale are represented by five items each; seven items are used for determining Externalizing Problems.

The PSC-17 was rated by preschool teachers from 12 elementary schools/child development centers in South Carolina that were involved in a federal grant project to provide information about young children's behavioral risk upon entry to school. A total of 1,000 preschool-aged children's PSC-17 ratings were obtained. Responses to items were provided for each student within a preschool classroom using a three-point frequency scale with anchors: 0 = "Never", 1 = "Sometimes", or 2 = "Often" based on occurrence of the listed behavior over the past several weeks. The Externalizing Problems subscale was used as this subscale was noted by teachers to be the area which teachers report as most problematic to the classroom environment (Greer, Wilson, DiStefano, & Liu, 2012). The PSC-17 Externalizing Problems are reported in Appendix A. Winsteps (version 4.4.1; Linacre, 2019) was used for all Rasch RSM analyses.

To assess unidimensionality of the Externalizing Problems subscale, an unrotated PCA of standardized residuals and the standardized residual contrast plot were examined. This analysis is used to determine if there is additional variance to be explained after the latent construct has been extracted (Linacre, 1992). As recommended, the construct should account for at least 50% of the total variance to be explained and,

Table 3.1 Category frequencies and average measures for PSC-17 screener, Externalizing Problems subscale

Category label	Observed count	Average measure[a]	Infit MNSQ	Outfit MNSQ	Threshold
0—Never	4884	−2.69	1.01	1.01	None
1—Sometimes	1654	−0.85	0.97	0.90	−1.37
2—Often	454	0.99	1.05	1.12	1.37

[a]Average Measure = sum (person measures—item difficulties)/count of observations in category

after accounting for the model, remaining extracted components should account for a small percentage of the remaining variance (less than 5%; Linacre, 1992). The PCA of the standardized residuals showed that the dimension extracted by the Rasch model account for 47.8% of the variance by the persons and items, slightly lower than recommendations. In addition, the unexplained variance in the first extracted component was 11.7% which was higher than the recommended value of 5%. Part of the reason for the high level of unexplained variance was thought to be due to the small number of items on the Externalizing Subscale. Overall, the results showed that the Externalizing Problems subscale shows some characteristics of dimensionality; however, we recognize that this assumption tentatively holds, allowing this subscale to be used to illustrate the Rasch RSM.

Externalizing Subscale: Category Usage. Table 3.1 showed the example output for the three-category rating scale. As we can see that all three category frequencies were larger than 10 responses, and the distribution of responses per category was right-skewed. The right-skewness in this situation shows that most of the students are not demonstrating externalizing problems.

The average measure for category 0 was −2.69 logits, and increased monotonically, moving from category 1 (: at −0.85 logits), to category 2 at 0.99 logits. It was expected that the higher the category selected, the higher the student's average measures. Category Infit and Outfit results were within the acceptable range. Thresholds results illustrated that the PSC-17 rating scale met the criteria that thresholds should increase by at least 1.4 logits between categories but not more than 5 logits (Linacre, 1999).

Externalizing Subscale: Response Probabilities and Thresholds. The graph in Fig. 3.3 illustrates the probability of responding to each category, given the difference in estimates between person ability and any item difficulty (Bond & Fox, 2007). As noted, each category has a definite peak, showing it is the most probable response for teachers at least some of the time. The threshold estimates were identified in Fig. 3.3 by dashed lines between curves. For ratings of 0, 1, 2, the threshold estimates were −1.37 and 1.37, respectively. In sum, this information suggests that the 0–2 rating scale is functioning appropriately.

Externalizing Subscale: Wright Map. Calibrated scores for both children and items are provided in the Wright map shown in Fig. 3.4. On the left side of the Wright map are the person measures, showing the placement of children by their estimated

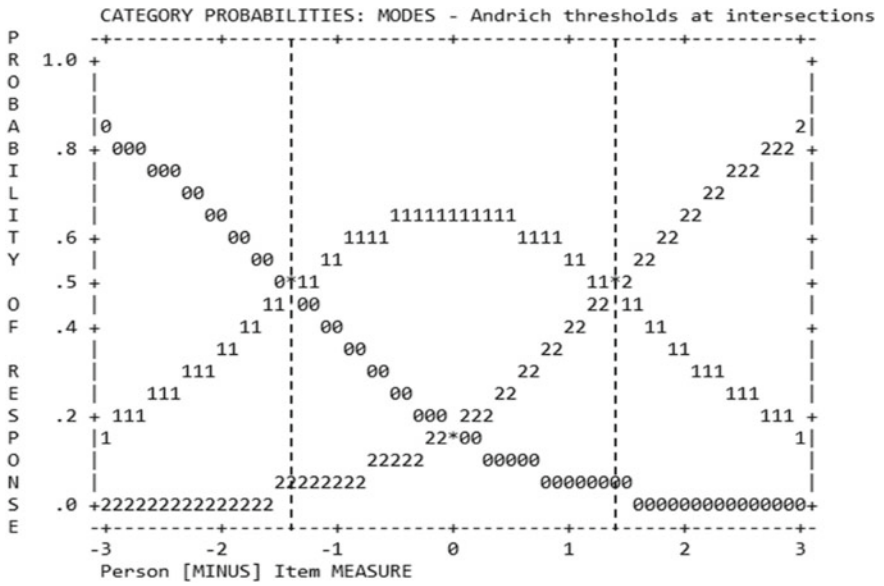


Fig. 3.3 Response category probability curves for the Externalizing Problems subscale of PSC-17

"externalizing problem" scores, and information about the relationship between items and construct is presented on the right side. Both person measures and item measures are on the same scale which children's latent scores can be interpreted related to the placement of the items. For person and item distribution of scores, the mean (*M*) of distribution is noted, with one (*S*) and two (*T*) standard deviations from the mean noted.

The left side of the graph provides information about the distribution of children rated by teachers. As we can see that most preschoolers were rated by teachers are relatively well-behaved—this is seen by the low average value of the person latent score (reported as −0.4) and the majority of children noted by a code of "X" or "." (relative to the number of cases) at the lower end of the scale. On the right side, the PSC-17 Externalizing items can be compared to the distribution of child ratings. These items are used for identifying a range of severe externalizing problems included on the screener. Items at the top of the item distribution are more sever and harder for teacher to frequently observe in the classroom, and items at the bottom of the scale (i.e., "Fights with other children" and "Does not listen to rules") are easier for teachers to observe. These two items are between 1 (*S*) and 2 (*T*) standard deviations below the item measure mean. Also, the three items at the top of the Wright map at the same "line" are not providing unique information regarding externalizing problems in young children. These items all are at roughly 1 standard deviation above the item mean; future revisions of the PSC-17 Externalizing Problems subscale may want to consider incorporating different items that help to identify children along the latent continuum.

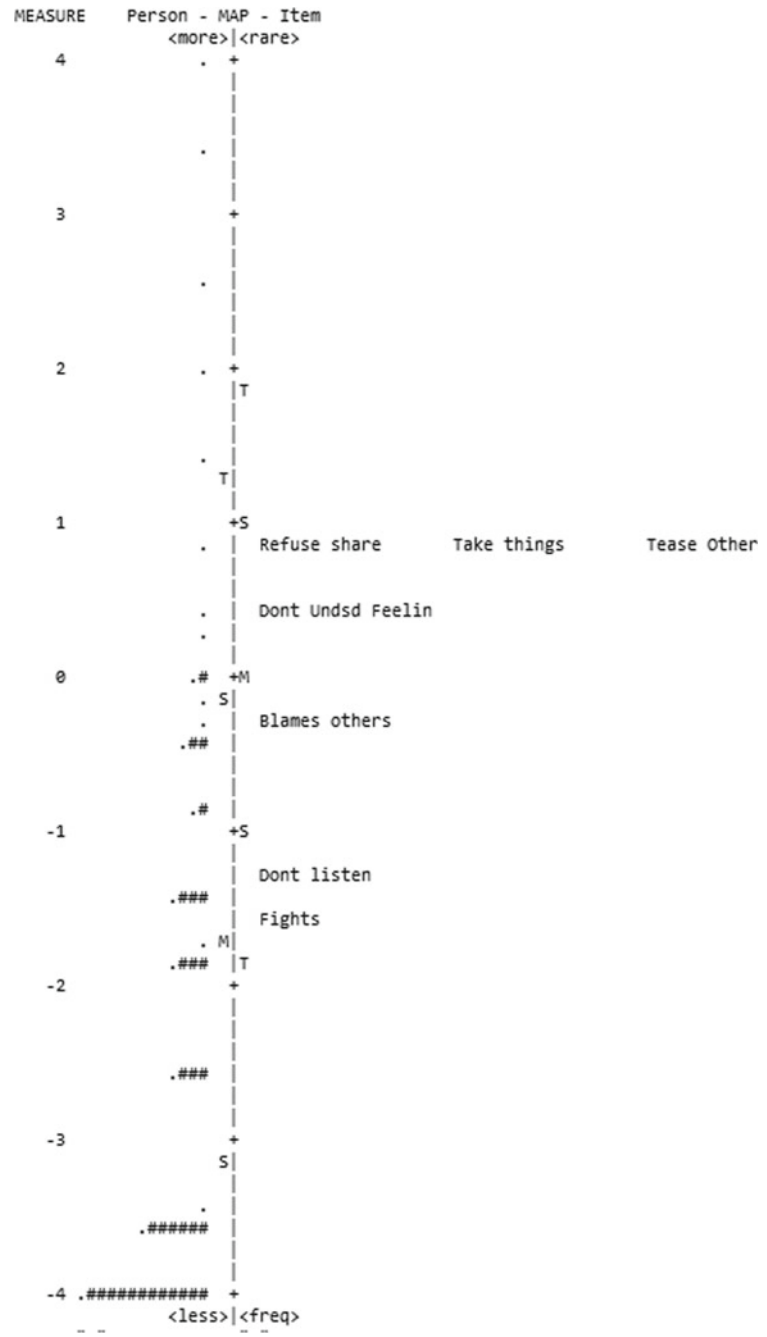


Fig. 3.4 Wright map for the PSC-17 Externalizing Problems subscale

Figure 3.5 presents the Wright map when there are ordinal scales. The right-hand column shows the items positioned at the measures where the expected score on the item is equal to the category number. It is also the measure at which the category has the highest probability. The left-hand column shows the distribution of person ability measures along the variable. As we can see, children with low externalizing problems

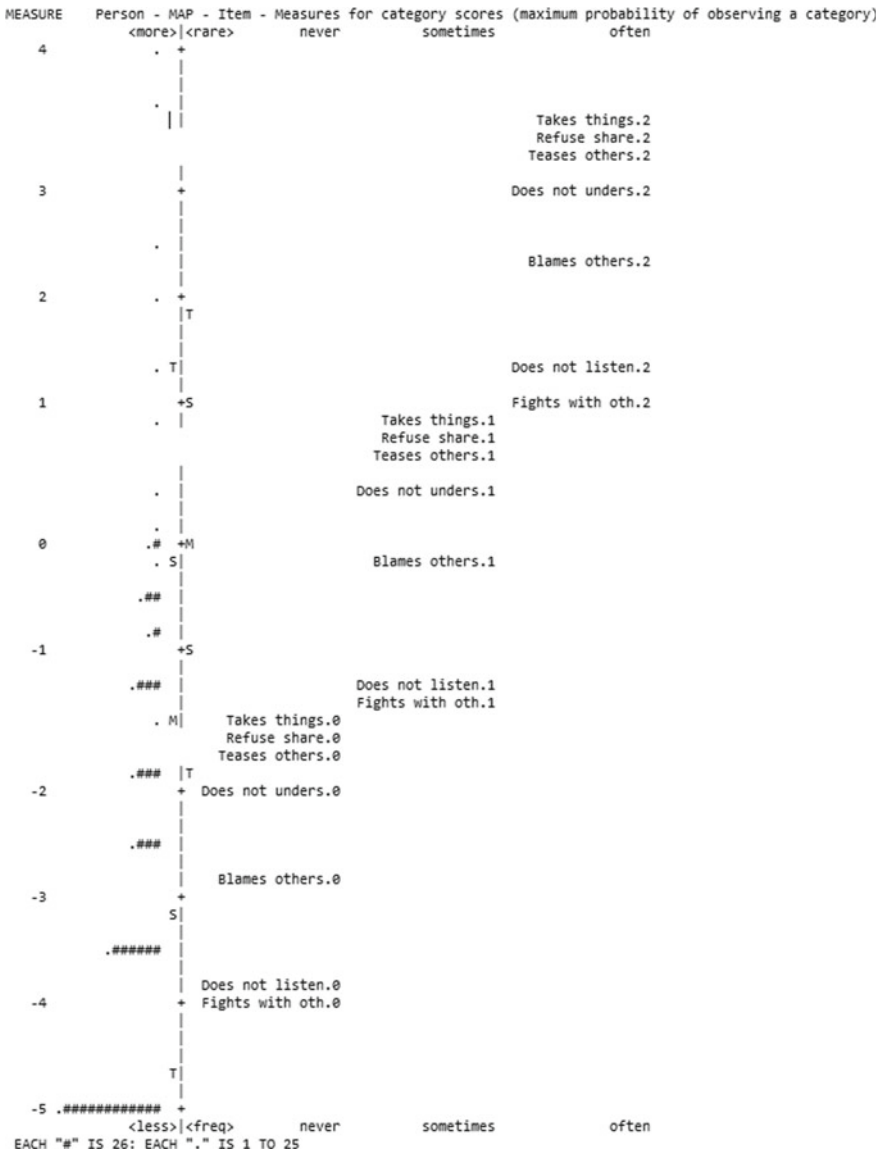


Fig. 3.5 Wright map measures by category scores, PSC-17 Externalizing Problems subscale

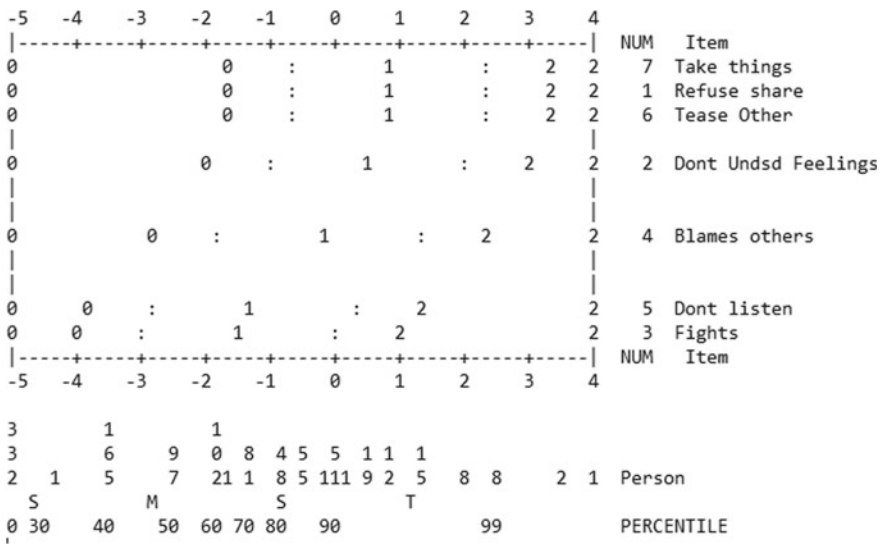


Fig. 3.6 Expected scores on the "Externalizing Problems subscale" of PSC-17 by child measure

(at the bottom of the Wright map) are likely to be rated zero by teachers using PSC-7 Externalizing Problem items; children with moderate externalizing problems (in the middle of the Wright map) are likely to be rated one by teachers; and children with high externalizing problems (on the top of the Wright map) are likely to be rated two by teachers using the seven PSC-17 Externalizing Problems items. By looking at the ordering of the item categories and the children's measures, we can conclude that all the items perform well and match what they are intended to be measured.

Figure 3.6 presents the expected item endorsements for children at various risk levels. Along the x-axis, preschoolers' risk levels are shown; along the y-axis are items from the PSC-17, ordered according to item difficulty values. Values correspond to the rating scale 0, 1, and 2, and colons correspond to threshold values, where a teacher would mark the next highest category on the rating scale if the threshold is surpassed. The response scales are approximately of equal distance apart, showing that the responses are spread among the different categories. Also, the response scale categories display a logical ordering of values (e.g., 0:1:2), illustrating that the categories are being used appropriately.

Determining Between Using the Rasch RSM and PCM

The Rasch RSM is not the only method available for analyzing ordinal data. The Rasch Partial Credit Model (PCM; Wright & Masters, 1982) is another option that researchers may consider. The PCM is similar in the sense that it accommodates

ordinal by including threshold values in its estimation. However, there are distinct differences between the RSM and PCM. RSM is typically used when all items on a questionnaire follow the same response scale (e.g., all items employ a 5-category Likert scale). PCM can be used in situations where the response scale differs across the questionnaire. Thus, each item is thought to have a unique rating scale. By allowing the items to have unique rating scales, the number of parameters to be estimated with the PCM increases by $(L-1) * (m-2)$, where L is the number items and m the number of categories in the rating scale (Linacre, 2000). While increasing the number of parameters may help to reduce misfit, generally, fewer rating scale parameters is preferred for stability and the communication of results.

To determine between use of RSM or PCM, Linacre (2000) recommends the following steps. First, examine the number of responses per category with the PCM. If there are categories with fewer than recommended responses (i.e., <10 ratings), estimates of difficulty of the parameters may be compromised. Second, communication of the results is facilitated if all items (or groups of items) share the same response format, (e.g., Strongly Disagree, Disagree, Agree, Strongly Agree). In such situations, the questionnaire/test developer and the respondents generally perceive the set of items to share the same rating scale. To attempt to explain a separate parameterization for of each item would hinder communication of the results. If there are only a few items that have a different scale (e.g., True/False), it may be easier to omit the non-conforming items than to argue that a separate scale exists for every item.

Future Directions. As with other areas of measurement, there are many unanswered questions which may be investigated using the Rasch RSM. For example, guidelines exist about the number of cases needed for stable estimation, including roughly 10 cases per category. An interesting avenue of investigation would be to examine the differences in estimated parameters with different numbers of sample sizes to determine how the minimum requirements change when scale usage follows patterns that may be observed with empirical studies, such as negatively and positively worded items on the same scale, respondents using the end points or the middle category of a Likert scale and investigation of parameter bias when items are skewed in opposite directions.

In addition, various software packages (e.g., IRTPRO, Xcalibre, the R-extended Rasch modeling package [eRm]; WinGen, Stata) are available to run the Rasch RSM. Differences among packages, including fit information and estimated parameters may be of interest to researchers. Such evaluations would not only compare results across software packages, but allow a thorough investigation of the drawbacks, benefits, and unique features offered by different software packages and programs. Finally, it may be of interest for researchers to include validity studies as part of the support for scaling decisions made from Rasch RSM. For example, examining relations between person-measure scores and relevant outcomes may provide quantitative data to support deleting misfitting items, changing the number of scale responses, and eliminating items which do not provide unique information to a scale.

In summary, the Rasch RSM is a useful model to use to examine characteristics of questionnaire data and for use in scale development. The methodology provides

an opportunity for researchers to investigate category usage, distributions of person-item measures for a scale, and estimate responses given characteristics of a person and item. In addition, visual representations of these procedures aid researchers and help to convey complex information with ease. We hope that this chapter will help to encourage more applied researchers to consider incorporating the Rasch RSM as part of their own investigations with questionnaire data.

Acknowledgements The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305A150152 to South Carolina Research Foundation. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

Appendix: PSC-17 Externalizing Problem Subscale Items (Gardner et al., 1999)

1. Refuses to share
2. Does not understand other people's feelings.
3. Fights with other children.
4. Blames others for his or her troubles.
5. Does not listen to rules.
6. Teases others
7. Takes things that do not belong to him or her.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ.

DiStefano, C., Pate, R., McIver, K., Dowda, M., Beets, M., & Murrie, D. (2016). Creating a physical activity self-report form for youth using Rasch methodology. *Journal of Applied Measurement*, *17*(2), 125–141.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11.

Fink, A. (2012). *How to ask survey questions* (Vol. 4). Thousand Oaks, CA: Sage Publishers.

Finney, S., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Greenwich, CT: Information Age.

Fowler, F. J. (2013). *Survey research methods*. Thousand Oaks, CA: Sage Publishers.

Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., …, & Chiappetta, L. (1999). The PSC-17: A brief pediatric symptom checklist with psychosocial problem subscales. A report from PROS and ASPN. *Ambulatory Child Health*, *5*, 225.

Greer, F. W., Wilson, B. S., DiStefano, C., & Liu, J. (2012, December). Considering social validity in the context of emotional and behavioral screening. In *School psychology forum* (Vol. 6, No. 4).

Iramaneerat, C. H. E. R. D. S. A. K., Smith E. V., Jr., & Smith, R. M. (2008). An introduction to Rasch measurement. *Best Practices in Quantitative Methods*, 50–70.

Jellinek, M. S., Murphy, J. M., Robinson, J., Feins, A., Lamb, S., & Fenton, T. (1988). Pediatric symptom checklist: Screening school-age children for psychosocial dysfunction. *The Journal of Pediatrics, 112*(2), 201–209.

Kahler, C. W., Strong, D. R., & Read, J. P. (2005). Toward efficient and comprehensive measurement of the alcohol problems continuum in college students: The brief young adult alcohol consequences questionnaire. *Alcoholism: Clinical and Experimental Research, 29*(7), 1180–1189. http://dx.doi.org/10.1097/01.ALC.0000171940.95813.A5.

Linacre, J. M. (1992). *Many facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3,* 103–122.

Linacre, J. M. (2000). Comparing and choosing between "partial credit models" (PCM) and "rating scale models" (RSM). *Rasch Measurement Transactions, 14*(3), 768.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. Rasch measurement: The dichotomous model. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2019). *Winsteps® (Version 4.4.1) [Computer Software]*. Beaverton, Oregon, Winsteps.com. Retrieved January 1, 2019, from https://www.winsteps.com/.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156.

Nardi, P. M. (2018). *Doing survey research: A guide to quantitative methods* (4th ed.). Philadelphia: Routledge.

Sick, J. (2010). Assumptions and requirements of Rasch measurement. *JALT Testing & Evaluation SIG Newsletter, 14*(2), 23–29.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*(1), 33.

Smith, E. V., Jr., Wakely, M. B., De Kruif, R. E., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement, 63*(3), 369–391.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18*(3), 291–307. https://doi.org/10.1177/1073191110374797.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.