Cakrawala Pendidikan
Jurnal Ilmiah Pendidikan
Vol. 41 No. 1, Februari 2022, pp.97-111
https://journal.uny.ac.id/index.php/cp/issue/view/2208
DOI: https://doi.org/10.21831/cp.v41i1.39130

Applying the Rasch Rating Scale Model (RSM) to investigate the rating scales function in survey research instrument

Jimmy Chong, Siti Eshah Mokshein*, Ramlee Mustapha
Universiti Pendidikan Sultan Idris, Malaysia
*Corresponding Author: eshah@fpm.upsi.edu.my

ABSTRACT

The functionality and quality structure of a rating scale are vital in a survey instrument. This article discusses the underlying assumption of a rating scale and the application of the Rasch Rating Scale Model (RSM) to diagnose the rating scale structure in a survey instrument. The instrument used to demonstrate the process of diagnosis a rating scale functioning in this study was the Vocational Teachers' Assessment Literacy (VoTAL) instrument. The VoTAL instrument used the five-point Likert-type response format and consisted of 88 items representing three assessment literacy constructs. The data were obtained from randomly selected 224 vocational teachers at five vocational colleges from Selangor State and the federal territory of Kuala Lumpur in Malaysia. The rating scale diagnosis results revealed that the initial five-step of rating scale categories were not effectively functioning as intended. It showed that the Andrich threshold of category three was disordered as it was not monotonically rich with categories. In addition, the result also found that the width of the threshold between category three and category four was too narrow (0.66 logits). The results indicate that the rating scale used in the VoTAL instrument was disordered. Thus, the initial five rating step categories collapsed down into four categories. In conclusion, this study highlights the importance of assessing the rating scale's functionality in a survey instrument to reduce measurement error and collect valid and reliable data.

Keywords: rasch measurement model, rating scales, rating scales function, survey instrument

Article history

Received:	*Revised:*	*Accepted:*	*Published:*
19 November 2021	*22 December 2021*	*15 January 2022*	*19 February 2022*

Citation (APA Style): Chong, J., Mokshein, S. E., Mustapha, R. (2022). Applying the Rasch Rating Scale Model (RSM) to investigate the rating scales function in survey research instrument. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan*, *41*(1), 97-111. https://doi.org/10.21831/cp.v41i1.39130

INTRODUCTION

Surveys in social science research remain one of the most popular data collection methods (Bradley, Peabody, Akers, & Knutson, 2015). Survey instruments are easy to administer in various ways, such as mail, online, paper-pencil, telephone, etc. With substantial studies utilizing survey design in social research, it is becoming increasingly essential that the instruments function as expected and measure what is supposed to measure. The instrument's precision and quality play crucial roles in ensuring the collected data are reliable and valid. Hence, to identify the problems that may distort the collected data's reliability and validity, it is vital to confirm the instrument's quality at the early stages of the instrument development (Bond & Fox, 2015).

The typical approach and the most popular type of instrument used in survey studies that have been around for a long is the rating scale. Therefore, there has been a significant number of research devoted to the rating scale functioning, especially the number of steps categories used in a rating scale (DiStefano & Jiang, 2020; Preston & Colman, 2000). The research into the rating scale steps categories has been becoming a debate for decades, especially on a Likert-type scale. Categorizations in the rating scale should be univocal, mutually exclusive, exhaustive, and well-defined (Linacre,

2002). In 1932, Rensis Likert made an early finding that differential category weighting structure was unproductive and beyond ordinal numbering. Rensis Likert later introduced the five-point category agreement rating scale. In 1967, Nunnally introduced an even category by removing the middle one in rating scales such as the Likert scale. In terms of Psychometric theory, Nunnally (1967) also reported that using more categories has more advantages than fewer categories. However, he also agrees that plentiful categories will confuse and irritate studies respondents. In addition, a study by Stone and Wright (1994) and Zhu, Updyke, and Lewandowski (1997), by changing five ordered categories into three ordered categories, found that fewer categories produce higher test reliability. Then, Lei Chang (1994) compares the output of four and six categories scales, and he reported that the four categories scale produced higher test reliability. In contrast, other scholars (e.g., Finn, Ben-Porath, & Tellegen, 2015; Hilbert, Küchenhoff, Sarubin, Nakagawa, & Bühner, 2016; Weng, 2004) had reported that a higher number of step categories would produce higher test reliability. Therefore, although the rating scale is customary in social research, there has been no consensus concerning the appropriate number of criteria used in the rating scale.

If researchers are still uncertain about the appropriate number of categories in the rating scale, examining the rating scale structure and functioning is always worthwhile (Linacre, 2002). In improving and verifying the functionality of rating scale categorization, Rasch measurement (Rasch, 1960) provides a practical method for rating scale analysis. Therefore, the purpose of this study was to demonstrate the processes to diagnose the rating scale's quality structure in a survey questionnaire by employing the Rasch measurement model, particularly the Rasch Rating Scale Model (RSM).

Rating Scales As Instruments In Social Research

The rating scales that Rensis Likert (1932) first established are widely used in the education and social sciences to assess latent variables. Survey instruments with ordered-categories rating scales are prevalent in education and social sciences research. Rating scales are employed when there is no instrument available, such as the natural sciences' measurement tools to measure latent constructs in the social sciences (Andrich, 2011). In social sciences, the Likert-type scale is the most commonly used tool for data collection. The Likert-type scale enables researchers to collect various data on perspectives (e.g., likelihood, agreement intensity, etc.) that make it highly versatile to most research conditions (Andrich, 2011; Fowler, 2013). Apart from that, the number of step categories on the Likert-type scale is always with the five-step categories. Likert-type scale is subject to change to provide more or fewer step categories, such as adding or removing the middle or creating a more continuous type of step categories (Fink, 2003; Fowler, 2013; Nardi, 2015). Referring to Smith, Wakely, De Kruif, and Swartz, (2003), the Likert-type scale provides three primary assumptions. At first, the Likert-type scale enables researchers to focus on the relevant research areas that need to study. Secondly, the Likert-type scale structure includes a range of potential responses to each item in a questionnaire. At last, all respondents are using the same stimuli to formulate their responses.

Regardless of the latent construct that the researchers are trying to assess, the use of the Likert-type scale may have several commonalities (Bond & Fox, 2015). Firstly, a statement serves for each item on the scale. Secondly, respondents are required to respond by choosing the degree of their level of agreement to a statement on the given rating scales. Lastly, the given scale for each statement is in the form of a rating scale with two or more step categories. The rating scale might be in the form of five steps agreement categories such as "strongly disagree" (1), "disagree" (2), "neutral" (3), "agree" (4), and "strongly agree" (5). The given rating scale might also may be in the form of four steps likelihood categories such as "very likely" (1), "unlikely" (2), "likely" (3), "very likely" (4). The odd number type of step categories may enable respondents to choose a "neutral" option for their response. In contrast, the even number type of steps categories leaves respondents with no choice but to accept either a negative or positive response (Bond & Fox, 2015).

Psychologists are generally finding a Likert-type scale the softer approach of data collection, as researchers explicitly maintain that the statements of items on the questionnaires are simply just to collect an opinion (Bond & Fox, 2015). Thus, the subjective nature of the Likert-type scale makes it easily be assigned to collect data that involves various forms of human conditions (Hales, 1986). However, even with the researchers' best effort in developing rating scales questionnaire, a variety of issues may still arise (Smith et al., 2003). Respondents might not use the rating scales as was expected. Respondents might not use the rating scales as was expected. Respondents may select socially appropriate responses, misunderstand the confusing items, or answer in a set of responses. Moreover, respondents' interpretation of the rating scale is solely based on their understanding of the rating labels. The use of ambiguous labels on a rating scale such as "sometimes" or "often" may cause distinctive use of the category. At the same time, the use of middle categories such as "neutral" or "unsure" may not be sharing the attribute measured by the other categories (Bradley et al., 2015; Smith et al., 2003). Additionally, the rating scale that only labeled the first and last points of the scale categories may leave respondents guessing and make their interpretation of the unlabelled responses categories. Further, the use of too many categories will bring more noise to the scale that could distress respondents to make an idiosyncratic selection of answers and may cause some categories to improperly function or not be used by respondents (Bond & Fox, 2015).

Therefore, the number of step categories in a questionnaire should be optimized by using a specific method to ensure that each category is effectively used to collect valid and reliable research data. Rasch Model, particularly the Rasch Rating Scale Model (RMS), is a reliable method used to diagnose the number of step categories in a questionnaire without the need to collect several sets of data from the distinct version of the same scale (Smith et al., 2003).

Rasch Rating Scale Model (RSM)

The conventional approach of analyzing rating scale data such as the Likert-type scale relies on the belief that all items in a questionnaire have the same level of difficulty, and the increased unit across each step of a category have equal value(Bond & Fox, 2015; Boone, 2020; DiStefano & Jiang, 2020). For example, from a set of five steps Likert scale, "strongly disagree" (1), "disagree" (2), "neutral" (3), "agree" (4), and "strongly agree" (5), the fifth (5) category has five times the value of the first category (1). This practice was used in the previous studies when researchers assumed the ordinal data of a rating scale as interval data by summing up the coded steps category in a questionnaire to create a total scale score (Boone, 2020). The calculated total score is then used in statistics to compare respondents. As stressed by Boone (2020), the rating scale, such as the Likert-type scale, cannot be assumed to have equal units across each step category, and each item on a questionnaire does not have the same level of difficulty. This traditional approach of analyzing rating scale data has ignored the subjectivity of the data by making unjustified assumptions about the nature of the scale (Bond & Fox, 2015; Boone, Staver, & Yale, 2014). Hence, this practice may lead to invalid mathematical operation and ineffective treatment of statistical analysis (Merbitz, Morris, & Grip, 1989; Wright & Linacre, 1989; Smith et al., 2003; Mokshein, Ishak, & Ahmad, 2019)).

However, the Rasch model treatment of rating scales data, particularly RSM, does not assert the conventional belief. In the Rasch model, items are categorized based on the item difficulty level. Each item in a questionnaire has its level of difficulty. The level of difficulty for each item is different. In a survey questionnaire, one particular item might be easier to agree on, and the other particular item might be harder to agree on. For example, in a five items questionnaire, the difficulty of respondents to agree with item one is not the same as the difficulty to agree with item five. In addition, Rasch RSM allocates each step of the category based on realistic nature (Bond & Fox, 2015). For example, on a Likert scale, Rasch RSM recognizes the values increased from "strongly disagree" (1) to "disagree" (2) does not have the same values increased from "agree" (4) to "strongly agree" (5). The Likert scale label of "strongly disagree"; "disagree"; "neutral"; "agree"; and "strongly agree" with

 e-ISSN 2442-8620, p-ISSN 0216-1370

the following coding of 1, 2, 3, 4, and 5 is only used to acknowledge the ordered step categories, not to be summoned to create a total score (Bond & Fox, 2015; Boone, 2020). As highlighted by Bond and Fox (2015), and Boone (2020), data collected with the rating scale has always been ordinal, and the number is the order of the rating scale.

The Rasch RSM developed by David Andrich (1978) is the extended version of the Rasch dichotomous model. RSM is specifically used to analyze ordinal data in a rating scale such as the Likert-type scale with a fixed number of step categories in a questionnaire (Wright & Masters, 1982; Linacre, 2000; Dimitrov, 2014; Engelhard & Wind, 2017). RSM analyzed the ordinal data on a rating scale by estimating the person measures (ability) and item location (difficulty) values on a single interval measurement scale (Mokshein et al., 2019). The unit values are known as logits, referred to as the logarithm of odds. Besides, RSM also estimates the thresholds value that signifies the width between each step category on a rating scale. The threshold on the rating scale is the point where a step from one to the nearby category. The number of entries on a rating scale is determined by the number of step categories minus one (DiStefano & Jiang, 2020). For example, the number of thresholds on a five-point Likert scale would be four, and the number of entries on a four-point rating scale would be three. Each entry on the rating scale would have its difficulty estimate (Abd-El-Fattah, 2015). Therefore, the way of rating scale managed by Rasch RSM is mathematically more reasonable and naturally appropriate than the conventional belief (Bond & Fox, 2015). Hence, Rasch RSM is the solution to convert the ordinal nature of rating scale data into interval data based on actual statistical proof.

Hence, the RSM equation can be expressed as below:

$$P(X_{ni} = x) = \frac{exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{x=0}^{m} exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}, x = 0,1,\ldots,m \quad (1)$$

This formula contains three main parameters as follow:
β_n is the measure (ability) of person *n*,
δ_i is the difficulty of item *i*,
τ_i is the threshold parameters
Where $P\ (X_{ni} = x)$ is the probability that a person *n* is observed in the rating scale category *x* on item *i*, which has $m + 1$ rating scale categories, and

$$\sum_{j=0}^{0} [\beta_n - (\delta_i + \tau_j)] = 0$$

From the formula (1), the Rasch RSM can be specified as the probability of a person *n* with the ability of β_n, is observed in a particular step category *x* on item *i* with the difficulty of δ_i (Andrich, 1978; Smith et al., 2003). Thus, every rating scale for a survey questionnaire should be analyzed to optimize the scale category effectiveness and to ensure that each category is appropriately functioning. For this purpose, the Rasch RSM provides the most convenient and simplest model to validate and improve the rating scale categorization functioning to increase the measurement accuracy (Linacre, 2002).

METHOD

Instrumentation

This study employed the Rasch RSM to analyze the rating scale categories of a measurement instrument. The instrument used to demonstrate the process of diagnosis a rating scale functioning in this study was the Vocational Teachers' Assessment Literacy (VoTAL) instrument. The VoTAL instrument was developed to assess vocational teachers' self-perceived assessment literacy. The VoTAL instrument consists of 88 items statements that represent three assessment literacy constructs; (1) Assessment Foundation, (2) Use of Assessment, and (3) Assessment Quality. The VoTAL instrument utilized the five-point Likert-type response format, which is (1) "strongly disagree," (2)

"disagree," (3) "less agree," (4) "agree," and (5) "strongly agree." This instrument had gone through a rigorous content validity process before being used in this study.

Samples and Procedure

The data were from the pilot study for the VoTAL instrument conducted on 224 vocational teachers at five vocational colleges in Selangor and Kuala Lumpur, Malaysia. The five vocational schools were chosen through a simple random sampling method. Of these samples, 128 (57.14%) were male, and 96 (42.86%) were female. All unit samples possess the same characteristics as they are all vocational teachers teaching vocational subjects in the government's vocational colleges, which use the same curriculum. The data were collected in November 2019 by on-site administration of the printed questionnaires to the study participants. Approvals for the data collection were obtained through the Malaysia Education Planning and Research Division (EPRD), Technical and Vocational Education Division, and the director of the selected vocational college.

Data Analysis

Rasch RSM is used to analyze various types of analysis, such as fit statistics, unidimensional diagnosis, differential item functioning, etc. However, this study only utilized the Rasch RSM to demonstrate the diagnosis of the rating scale functioning of the VoTAL instrument. Rasch RSM is pertinent for assessing the item difficulty of a rating scale with two or more ordered step categories (Smith, 2003). As previously described, Rasch RSM can be specified as the probability of a person n with the ability of β_n, is observed in a particular step category x on item i with the difficulty of δ_i (Smith, 2003; Andrich, 1978). The Rasch RSM probability formula is shown in formula (1) from the previous section. The Rasch RSM formula is implemented by Winsteps software (version 4.5.2), which was then used to perform the data analysis in this study. Winsteps is one of the many software packages that are used to run Rasch RSM analysis. Winsteps software can produce empirical evidence in discovering the ability of respondents to understand and discriminate the step categories as well as to inform the rating scale quality (Linacre, 2002).

Thus, in an attempt to diagnose and optimize the rating scale category effectiveness, each analysis concerning the rating scale functioning such as category observations and distributions, average measures, outfit MNSQ, Andrich threshold, and the width of step categories were rigorously analyzed. In achieving the desired results, Linacre (2002) had provided seven criteria as a guideline to assist researchers in conducting the analysis. However, as advised by Linacre (2002), not every criterion is appropriate for any specific rating scale analysis. Therefore, based on the review of Rasch rating scale analysis in the literature along with the suggestion by Smith et al. (2003) and Dimitrov (2014), it is recommended that the following five criteria are applicable for most rating scale conditions:

1. Each category should have a minimum of ten observations,
2. Observed average monotonically progress with categories,
3. The value of the Outfit means square (MNSQ) is lower than 2.0,
4. Step calibration (Andrich threshold) monotonically progress with categories, and
5. Step calibration (Andrich threshold) advances by a minimum of 1.4 logits and a maximum of 5.0 logits.

These criteria were used as a guideline to analyze and optimize the function of rating scale categories in this study.

FINDINGS AND DISCUSSIONS

Findings

The rating scale functioning analysis is assessed based on the output table produced by Winsteps. The initial application of the VoTAL instrument shows that the instrument's reliability was high. The person and item reliability were .96 and .99, respectively. The rating scales analysis begins by investigating the instrument's category structure, as shown on the Winsteps output in Figure 1.

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
1	1	672	4	-.48	-.56	1.11	1.40	NONE	(-2.78)	1
2	2	2106	13	-.16	-.19	1.04	1.10	-1.52	-1.16	2
3	3	2172	13	.23	.31	.92	.98	.02	-.19	3
4	4	7695	47	.96	.97	.85	.80	-.64	1.05	4
5	5	3707	23	1.95	1.91	1.09	1.02	2.14	(3.29)	5

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

Figure 1. Category structure of VoTAL instrument

The result from Figure 1 was based on the previously mentioned criteria for optimization of rating scale category functioning as below:

Each category should have a minimum of ten observations: This criterion revealed as the observed count for each category ranged from 672 to 7695 observations. The lowest observations were in category one, and the highest observations were in category four. This frequency of observations was more than enough to estimate a stable rating scale structure.

Observed average monotonically progress with categories: The observed average was in ordered and consistently increased monotonically (-0.48 < -0.16 < 0.23 < 0.96 < 1.95) across the step categories. This result implies that, in general, respondents with lower ability progressively endorsed lower step categories, while respondents with higher ability progressively endorsed higher step categories. Thus, this criterion was fulfilled.

The value of the Outfit means square (MNSQ) is lower than 2.0: The result in Figure 1 shows that the outfit MNSQ values ranged from 0.80 to 1.40, indicating that the data set introduced more information with lower unexplained noise. Therefore, this criterion was also fulfilled.

Step calibration (Andrich threshold) monotonically progress with categories: This criterion was not met as the Andrich threshold values reported in Figure 1 were not monotonically progress with categories. The value of category three (-.64) was smaller than category two (.02), which leads to disordering thresholds. The disordering thresholds can also be observed graphically from the category probabilities response curves in Figure 2. The x-axis in Figure 2 is the respondents' measures relative to item difficulty of endorsement, and the y-axis is the probability of a particular category observed in logit values. From Figure 2, it is noted that the 'hilltops' of category three were missing and did not peak.

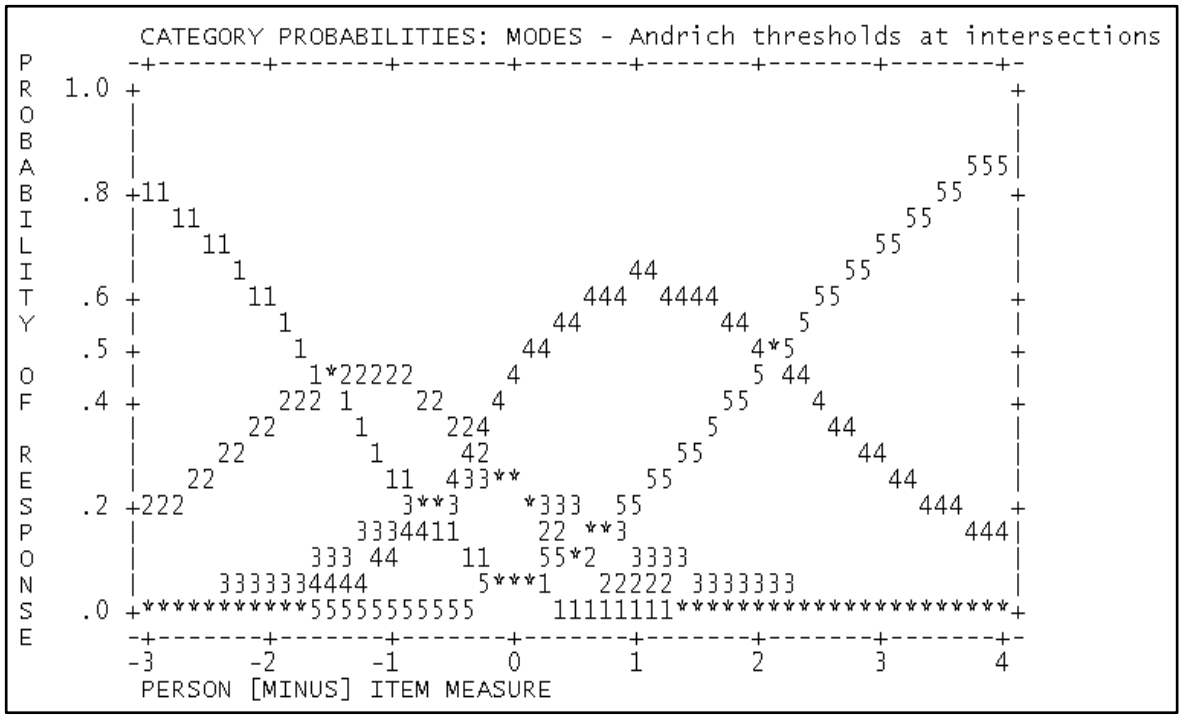

Figure 2. Category probabilities for VoTAL instrument

Step calibration advances by a minimum of 1.4 logits and a maximum of 5.0 logits: The Andrich threshold values across the step categories in Figure 2 did not meet this criterion. The width between each category threshold is given in Table 1.

Table 1. Thresholds width between categories

Thresholds between category	Thresholds width (logits)
Category 1 and 2	1.52
Category 2 and 3	1.54
Category 3 and 4	**0.66**
Category 4 and 5	2.78

From Table 1, it is found that the width between the Andrich threshold of category three and four was 0.66 logits (-0.64 – 0.02 = 0.66), which is less than the minimum requirement of 1.4 logits. This narrow threshold indicates that respondents might be unable to discriminate the categories. It may be a signal of categories overlapping. None of the entries advances more than 5.0 logits, as the maximum width of thresholds was 2.78 logits, which is the width of entries between categories four and five (-0.64 – 2.14 = 2.78).

The result from Figure 1, Figure 2, and Table 1 show empirical evidence that the rating scale categories of the VoTAL instrument were not properly functioning. Therefore, one possible remedy is those problematics categories should match the adjacent categories or be revised to improve the meaning and function of the rating scale (Linacre, 2002). From the above analysis, step category three was the most problematic. Thus, category three suggested being collapsed down into category two or collapsed up into category four. The result of both sets of collapse categories (category three collapse down into category two and collapse up into category four) was then compared to decide a better set of empirical categories. Collapsing down category three into category two and collapsing up category three into category four will result in the rating scale categories down to four categories instead of five. In the first set of collapsing categories (collapsing category three down into category two), the

first category remained as category one. The second and third categories were re-coded as category two. The fourth category was re-coded as category three. Lastly, the fifth category was re-coded as category four. The original code of "12345" in the Winsteps control file was then re-coded to "12234", and re-analyzed. In this case, the response of category three serves as the same response as category two. The summary of category structure and category probabilities curves resulting from the collapsing of category three down into category two is presented in Figures 3 and 4, respectively.

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
1	1	672	4	-.75	-.84	1.09	1.17	NONE	(-3.48)	1
2	2	4278	26	-.08	-.02	.93	1.00	-2.30	-1.22	2
3	3	7695	47	1.09	1.05	.83	.80	-.09	1.18	4
4	4	3707	23	2.31	2.34	1.12	1.10	2.39	(3.55)	5

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

Figure 3. Category structure for collapsing category three down into category two, "12234"

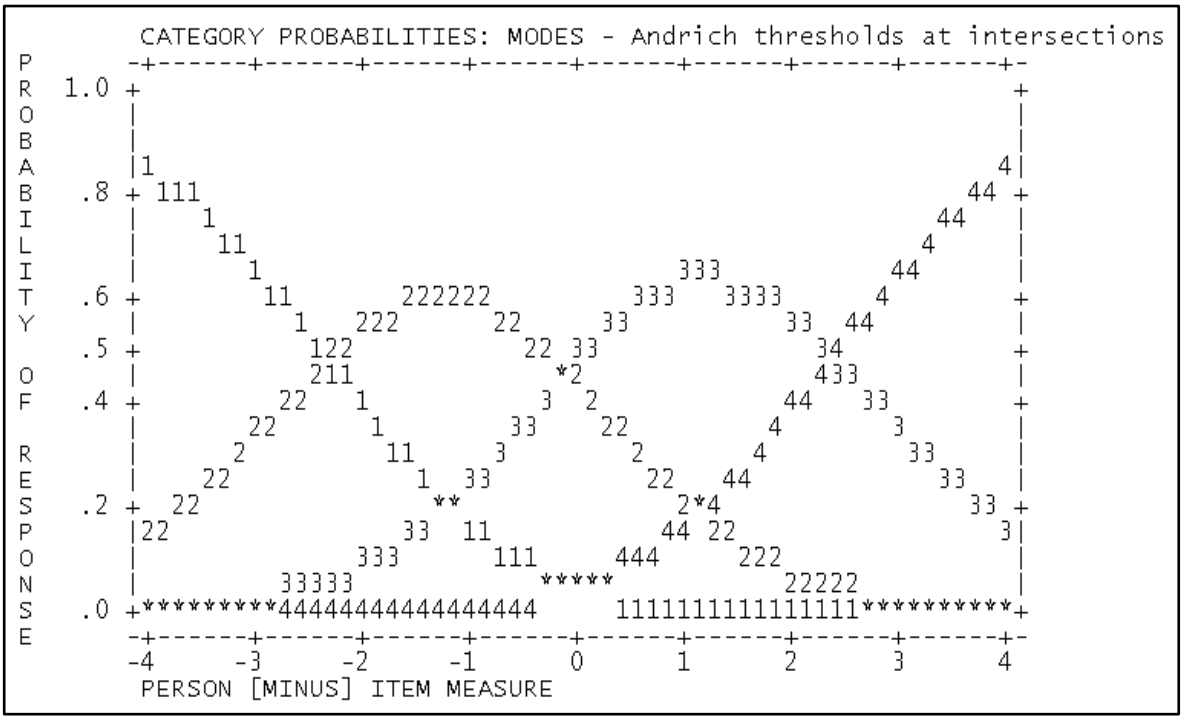

Figure 4. Category probabilities for collapsing category three down into category two, "12234"

In the second set of collapsing categories (collapsing category three up into category four), the first category remained as category one. The second category remained as category two. The third and fourth categories were re-coded as category three. Lastly, the fifth category was re-coded as category four. The original code of "12345" in the Winsteps command was then changed into "12334". The response of category three was analyzed as the same response as category four. The summary of category structure and category probabilities curves resulting from the collapsing of category three up into category four is shown in Figures 5 and 6, respectively.

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	OBSERVED SCORE	COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
1	1	672	4	-.62	-.78	1.11	1.25	NONE	(-3.00)	1
2	2	2106	13	-.19	-.08	.89	.87	-1.60	-1.37	2
3	3	9867	60	1.02	1.01	.90	.89	-1.11	.88	3
4	4	3707	23	2.54	2.54	1.06	1.00	2.71	(3.82)	5

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

Figure 5. Category probabilities for collapsing category three up into category four, "12334"

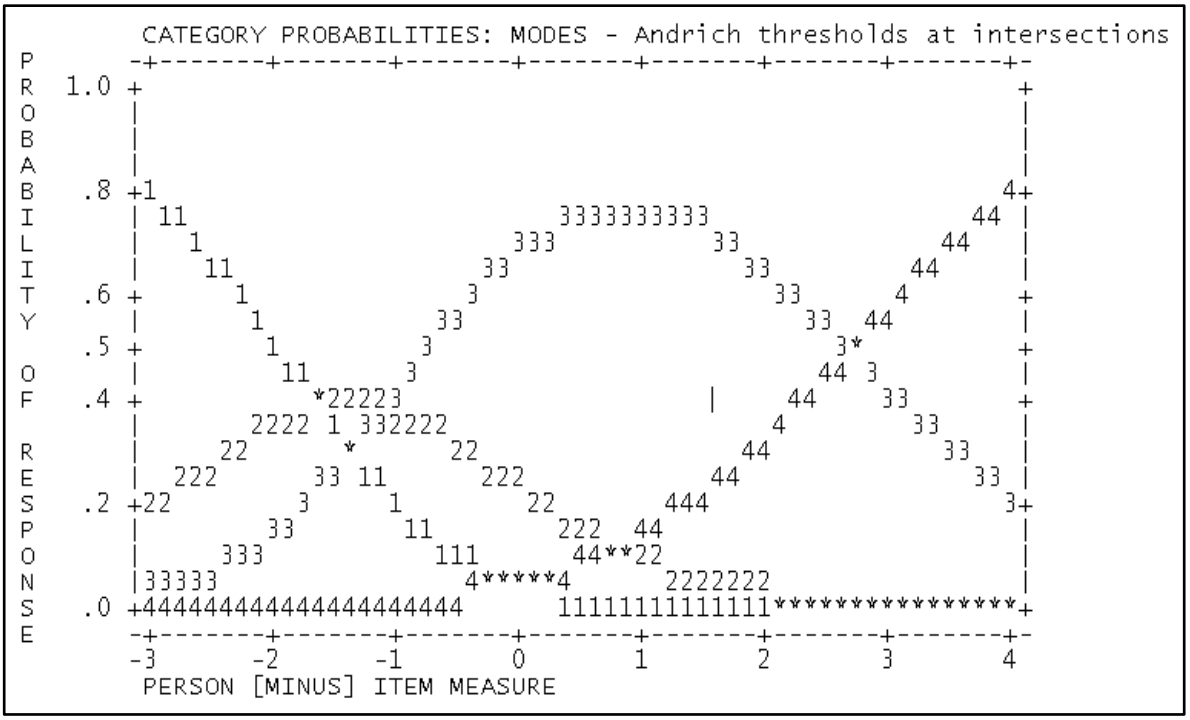

Figure 6. Category probabilities for collapsing category three up into category four, "12334"

In deciding which set of collapsing categories to be used, the comparison based on the guideline criteria was made. This comparison is seen in Table 3. The analysis in Table 3 clearly shows that collapsing category three down into category two "12234" fulfills all the criteria. However, collapsing category three up into category four "12334" has improved the order of the threshold but failed to meet the width of the minimum threshold of 1.4 logits. The thresholds width between category three and two was only 0.49 logits. The category probabilities curves resulting from the small width of thresholds are in Figure 6. As in Figure 6, although a "hilltop" can be seen in category two, the peak was unclear and almost sunken in-between categories one and three. In comparison to Figure 4, a clear peak can be observed for every category, indicating that all categories had a reasonable width of thresholds.

Table 2. Comparison of categorization of original "12345", collapsing category three down into category two, "12234", and up into category four, "12334"

Criteria	Original, "12345"	Collapsing category three down into two, "12234"	Collapsing category three up into four, "12334"
Minimum of 10 observations per category	✓	✓	✓
Observed average advance monotonically	✓	✓	✓
Outfit MNSQ < 2.0	✓	✓	✓
Andrich threshold advance monotonically	×	✓	✓
Andrich threshold advance by > 1.4 logits and < 5.0 logits	×	✓	×

Based on the comparison result, the final verdict of the analysis was to use the categorization of "12234". The collapse of category three to category two shows a better empirical rating scale improvement and can meet all the mentioned criteria. Hence, the labels of the new set of step categories were then re-arranged and re-coded as "strongly disagree" (1), "disagree" (2), "agree" (3), and "strongly agree" (4). The "less agree" label was eliminated. This new set of four-step rating scales will be tested again in the next data collection for final verification.

Discussions

A well-functioning rating scale used in an instrument is crucial in collecting valid and reliable data. Thus, this study demonstrates the method to analyze and optimize the rating scale function of the VoTAL instrument using the Rasch RSM method. At first, the VoTAL used five-step rating scale categories in data collection. The collected data were then analyzed using Winsteps software version 4.5.2, and the result was assessed based on the five criteria mentioned in the data analysis section. The initial five steps rating scales used in the VoTAL instrument were able to meet criteria one, two, and three but failed to fulfill criteria four and five.

For the first criteria, each category of the VoTAL instrument has more than ten observations. The result shows that the number of observations on each of the step categories was stable for step calibration estimations. The step calibration is determined by the log ratio of the number of observations of its nearby category. If the number of observations is low, the estimation of the step thresholds is imprecise and possibly unstable as it does not provide adequate responses for the estimation of the thresholds (Linacre, 2002). For that reason, a small change in the number of observations will shift the scale estimation structure. Thus, to get a stable rating scale, each step category should have a minimum of ten observations.

The second criterion was also fulfilled. The observed average was monotonically in line with categories. The observed average is the indicator that implies the use of each step category. Generally, higher measures should indicate higher categories, while lower measures should indicate lower categories. Thus, the observed average must be in order and monotonically advance with step categories. The disorder observed average could indicate the uncertain meaning of the rating scale that may lead to a controversial and imprecise measure (Linacre, 2002).

The third criterion required the Outfit means square (MNSQ) values to be lower than 2.0. This criterion was met, as from the result, it was reported that the highest Outfit MNSQ values were 1.40. The MNSQ is the ratio of chi-square statistics to its degrees of freedom(Linacre, 2019). Therefore,

the model specified a uniform value of randomness for MNSQ is expected to be1.0 (Wright & Panchapakesan, 1969). The Outfit MNSQ values of more than 1.50 imply that up to 50 percent or more unexplained noise in the data (Smith, 1996; Linacre, 2002; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). While Outfit MNSQ values exceed 2.0 implies that the unexplained noise is higher than the explained noise, showing that the responses data introduce more falsity than useful information (Smith, 1996; Linacre, 2002; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008; Mokshein et al., 2019). High Outfit MNSQ values (> 2.0) suggest that the data may not support meaningful measurement and associate with unexpected use of categories and noisy data.

The fourth criterion was not fulfilled. The requirement for the fourth criterion was the Andrich threshold should progress monotonically with step categories. Failure to comply with this criterion will lead to disordering thresholds. From the result, it was found that the Andrich threshold of category three was disordered. Disordering thresholds indicate low functionality of certain rating steps that result from the improper use of categories and irregular pattern of category frequency (Linacre, 2002). For example, a lower measure of persons was observed in the higher categories, and a higher measure of persons was observed in the lower categories. The disordered thresholds do not conform to the crucial conceptual aspect of the rating scale expected by the Rasch model, where a higher measure of persons should be observed in the higher categories and vice versa (Andrich, 1996; Linacre, 2002). In addition, the disordering threshold observed in Figure 2 implies that category three never be the highest probability to find at any point as the respondents' measures relative to item difficulty increase along the x-axis. Ideally, to shows that a particular category would be the most probability being endorsed at some point as the respondents' measure relative to item difficulty increase along the x-axis, each step category should peak and show clear 'hilltops' (Bond & Fox, 2015). Disordering thresholds, therefore, may degrade the quality of the data (Dimitrov, 2014).

The fifth criterion required the Andrich threshold advance by a minimum of 1.4 logits and a maximum of 5.0 logits. In this study, the analysis showed that the threshold width between category three and category four was too narrow, which is less than 1.4 logits. Thus, the fifth criterion was not fulfilled. The width between thresholds categories indicates that each step category specifies a specific location on the latent variable (Bond & Fox, 2015). Thus, to ensure the step calibration distinct location, the estimates threshold for each category on the logit scale should be neither too near nor too distant between them. A threshold width of fewer than 1.4 logits could be a signal of overlapping categories and may indicate that respondents might be unable to discriminate between the step categories (Bond & Fox, 2015; Dimitrov, 2014; Linacre, 2002). In contrast, a threshold width of more than 5.0 logits could deprive the precision of measurement in the center of a category, which will provide fewer details from the best-targeted respondents (Linacre, 2002). Hence, a narrow threshold width of fewer than 1.4 logits is suggested to be collapsed to the adjacent categories or redefined to have a more precise meaning (Linacre, 2002). On the other hand, a wider threshold width of more than 5.0 logits might be conceptualized or divided into two narrowed categories to allow the scale to be able to collect more comprehensive information (Linacre, 2002).

From the result in Figure 1, Figure 2, and Table 1, it was suggested that category three should either be collapsed down into category two or up into category four. The collapsing categories were done to reduce noise and improve the functionality of the rating scale (Wright & Linacre, 1992; Linacre, 1999). However, careful consideration is required in collapsing categories, and it needs to be aware of the qualitative meaning of each of the step categories (Bond & Fox, 2015; Smith et al., 2003). Collapsing categories should make sense. It would be compatible with collapsing "agree" and "strongly agree", but not for collapsing "agree" and "disagree" that has a different meaning. In this study, respondents may not be able to distinguish between category two, "not agree" and category three, "less agree" or category three, "less agree" and category four, "agree". The problematic category here was category three, which carried the label of "less agree". The "less agree" label was confusing. According to the Oxford Dictionary (2020), the meaning of "less" is "a smaller amount

 e-ISSN 2442-8620, p-ISSN 0216-1370

of". In this case, "less agree" means "a smaller amount of agreement". At the same time, some parts of "a smaller amount of agreement" may also carry the meaning of disagreement. Therefore, it is sensible to collapse the "less agree" category down into "disagree" or up into "agree".

The decision to choose either to collapse the "less agree" category (category three) down into "disagree" (category two) or up into "agree" (category four) was made by comparing the analysis results for both sets of collapsing categories to the required criteria. As previously shown in Table 2, collapsing category three down into category two was able to meet all the required criteria and demonstrate the functionality of all step categories. While collapsing category three up into category four failed to meet the width of the minimum threshold of 1.4 logits. Therefore, category three was decided to be collapsed down into category two. The initial five steps rating scale categories of the VoTAL instrument were re-coded to four steps rating scale categories with the new arranged labels of "strongly disagree" (1), "disagree" (2), "agree" (3), and "strongly agree" (4).

The decision to choose either to collapse the "less agree" category (category three) down into "disagree" (category two) or up into "agree" (category four) was made by comparing the analysis results for both sets of collapsing categories to the required criteria. As shown in Table 2, collapsing category three into category two enables it to meet all the required criteria and demonstrate the functionality of all step categories. While collapsing category three up into category four failed to meet the width of the minimum threshold of 1.4 logits. Therefore, category three was collapsed into category two. The initial five steps rating scale categories of the VoTAL instrument were re-coded to four steps rating scale categories with the new arranged labels of "strongly disagree" (1), "disagree" (2), "agree" (3), and "strongly agree" (4).

It is evident that collapsing problematics categories down or up into the adjacent categories would improve the function and meaning of a rating scale. Nevertheless, there was no assurance that the new set of categories would function as expected. Therefore, the diagnosis of the rating scale should be made at the pilot study stage. The new set of categories from the pilot data should be tested again in the next data collection to confirm the functionality of the new set of rating categories.

CONCLUSION

This study had systematically demonstrated the diagnosis of a rating scale functioning using Rasch RSM. The diagnosis of the rating scale was just one aspect of the Rasch RSM analysis. The instrument is subject to other types of analysis to assess the instrument's overall quality, such as item polarity analysis, person and item fit analysis, unidimensional diagnosis, Wright map analysis, differential item functioning (DIF), etc. From this study, it was suggested that anytime rating scales are used in an instrument, the rating step categories need to be empirically assessed before they can be claimed as appropriate for any study. This study also discusses the underlying assumptions of the rating scale in a survey questionnaire, reviews the parameter criteria used by the Rasch RSM in the diagnosis of the rating scale, and demonstrates the diagnosis process. Therefore, this study may benefit social science researchers in developing a survey questionnaire. Notably, this study provides a framework for assessing the functionality of the rating scale used in a survey questionnaire so that the data collected with properly functioning rating scales will be more precise, lower noise, and with a reduced measurement error.

REFERENCES

Abd-El-Fattah, S. M. (2015). Rasch Rating Scale Analysis of the Arabic Version of the Physical Activity Self-Efficacy Scale for Adolescents: A Social Cognitive Perspective. *Psychology*, *06*(16), 2161–2180. https://doi.org/10.4236/psych.2015.616213

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Andrich, D. (1996). Measurement Criteria for Choosing among Models with Graded Responses. In *Categorical Variables in Developmental Research*. https://doi.org/10.1016/b978-012724965-0/50004-3

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(5), 571–585. https://doi.org/10.1586/erp.11.59

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Third Edition*. Routledge.

Boone, W. J. (2020). Rasch Basics for the Novice. In M. S. Khine (Ed.), *Rasch Measurement* (pp. 9–30). Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3_2

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutson, N. M. (2015). Rating Scales in Survey Research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice*, *8*(1), 1–12. https://doi.org/10.29115/sp-2015-0001

Dimitrov, D. M. (2014). *Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields*. Wiley.

DiStefano, C., & Jiang, N. (2020). Applying the Rasch Rating Scale Method to Questionnaire Data. In M. S. Khine (Ed.), *Rasch Measurement* (pp. 31–46). Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3_3

Engelhard, G., & Wind, S. A. (2017). Invariant Measurement with Raters and Rating Scales. In *Invariant Measurement with Raters and Rating Scales*. https://doi.org/10.4324/9781315766829-14

Fink, A. (2003). *How to Ask Survey Questions* (2nd ed.). SAGE Publications, Inc.

Finn, J. A., Ben-Porath, Y. S., & Tellegen, A. (2015). Dichotomous versus polytomous response options in psychopathology assessment: Method or meaningful variance? *Psychological Assessment*, *27*(1), 184–193. https://doi.org/10.1037/pas0000044

Fowler, F. J. (2013). *Survey Research Methods*. SAGE Publications.

Hales, S. (1986). Rethinking the Business of Psychology. *Journal for the Theory of Social Behaviour*, *16*(1), 57–76. https://doi.org/10.1111/j.1468-5914.1986.tb00066.x

Hilbert, S., Küchenhoff, H., Sarubin, N., Nakagawa, T. T., & Bühner, M. (2016). The influence of the response format in a personality questionnaire: An analysis of a dichotomous, a likert-type, and a visual analogue scale. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *23*(1), 3–24. https://doi.org/10.4473/TPM23.1.1

Lei Chang. (1994). A Psychometric Evaluation of 4-Point and 6-Point Likert-Type Scales in Relation to Reliability and Validity. *Applied Psychological Measurement*, *18*(3), 205–215. https://doi.org/10.1177/014662169401800302

Likert, R. (1932). *A Technique for the Measurement of Attitudes* (Issue nos. 136-165). publisher not identified.

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*, 85–106.

Linacre, J. M. (2019). *A User's Guide to WINSTEPS MINITEP, Rasch-Model Computer Programs, Program Manual 4.4.7*. winsteps.com. https://doi.org/ISBN 0-941938-03-4

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, *70*(4), 308–312.

Mokshein, S., Ishak, H., & Ahmad, H. (2019). The Use Of Rasch Measurement Model In English Testing. *Jurnal Cakrawala Pendidikan*, 38(1), 16-32. doi:https://doi.org/10.21831/cp.v38i1.22750

Nardi, P. M. (2015). *Doing Survey Research*. Taylor & Francis.

Nunnally, J. C. (1967). *Psychometric theory*. McGraw-Hill.

Oxford Dictionary. (2020). /www.oxfordlearnersdictionaries.com/definition/english/less_1?q=less

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut.

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. https://doi.org/10.1186/1471-2288-8-33

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*(3), 516–517.

Smith, E. V, Wakely, M. B., De Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing Rating Scales for Self-Efficacy (and Other) Research. *Educational and Psychological Measurement*, *63*(3), 369–391. https://doi.org/10.1177/0013164403251320

Stone, M., & Wright, B. D. (1994). Maximizing rating scale information. *Rasch Measurement Transactions*, *8*(3), 386.

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956–972. https://doi.org/10.1177/0013164404268674

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however,

must be interval. *Archives of Physical Medicine and Rehabilitation, 70*(12), 857–860.

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

Wright, B. D., & Panchapakesan, N. (1969). A Procedure for Sample-Free Item Analysis. *Educational and Psychological Measurement*, *29*(1), 23–48. https://doi.org/10.1177/001316446902900102

Zhu, W., Updyke, W. F., & Lewandowski, C. (1997). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. *Journal of Outcome Measurement, 1*(4), 286–304.