ORIGINAL PAPER

Applying Rasch Methodology to Examine and Enhance Precision of the Baby Care Questionnaire

Quoc Cuong Truong [1] · Merideth Gattis[2] · Carol Cornsweet Barber[1] · Wendy Middlemiss[3] · Terry Au[4] · Oleg N. Medvedev[1]

Abstract

The Baby Care Questionnaire (BCQ) is an established ordinal measure of parenting beliefs about infant care, which includes structure and attunement scales. Rasch analysis is a powerful method to examine and improve psychometric properties of ordinal scales. This study aimed to evaluate the psychometric properties and improve precision of the structure and attunement scales of the BCQ using Rasch methodology. A Partial Credit Rasch model was applied to a sample of 450 mothers from the United Kingdom ($n = 225$) and New Zealand ($n = 225$) who completed the BCQ. Initial analyses indicated acceptable reliability of the structure and attunement scales of the BCQ, but some items showed misfit to the Rasch model associated with local dependency issues in both scales. After combining locally dependent items into testlets, both scales of the BCQ met expectations of the unidimensional Rasch model and demonstrated adequate and strong reliability and invariance across countries and person factors such as participants' age and their baby's sex. This permitted the generation of conversion algorithms to transform ordinal scores into interval data to enhance the precision of both scales of the BCQ. In conclusion, using Rasch methodology, this study demonstrated robust psychometric properties of the structure and attunement scales of the BCQ after minor modifications. The ordinal-to-interval conversion tables published here can be used to further enhance the precision of the structure and attunement scales of the BCQ without changing their original response format. These findings contribute to the enhancement of precision in measuring parenting beliefs about infant care.

Keywords Rasch analysis · Parenting beliefs · Structure · Attunement · Interval data.

Highlights

- The Baby Care Questionnaire (BCQ) assesses parenting beliefs on infant care through structure and attunement scales.
- Rasch methodology assessed BCQ's psychometric properties, refining structure and attunement scales for interval-level precision.
- By creating testlets, both BCQ scales demonstrated strong reliability and invariance across countries and person factors.
- Conversion algorithms converted ordinal scores to interval scores to enhance the precision of both scales of the BCQ.

Measurement of beliefs about parenting and what babies need is important because these beliefs, and how they change during the transition to parenthood, are an important window into understanding parenting behaviours, well-being, and their impact on child outcomes (Galbally et al., 2018; Hughes et al., 2012; Smetana, 2017; Tikotzky & Sadeh, 2009). The Baby Care Questionnaire (BCQ; Winstanley & Gattis, 2013) is a valid and reliable psychometric assessment tool to measure parenting beliefs about infant care across fundamental care domains. However, this parenting belief measure is an ordinal scale, which has limitations (Tennant & Conaghan, 2007). Technically, scores collected from ordinal measurement are inappropriate for statistical parametric testing because ordinal data cannot be

✉ Merideth Gattis
GattisM@cardiff.ac.uk

1 University of Waikato, School of Psychology, Hamilton, New Zealand

2 School of Psychology, Cardiff University, Cardiff, UK

3 University of North Texas, Denton, Texas, US

4 University of Hong Kong, Pok Fu Lam, Hong Kong

⚫ Springer

added, subtracted, multiplied or divided, so the calculation of means and standard deviations are invalid (Linacre, 2004; Merbitz et al., 1989; Norquist et al., 2004). Rasch methodology (Rasch, 1960, 1961) is a solution to enhance the psychometric properties of an ordinal scale to estimate an interval-level measurement if the strict assumptions of the unidimensional Rasch model are met (Hobart & Cano, 2009; Merkin et al., 2020; Tennant & Conaghan, 2007). This method can be used to develop ordinal-to-interval conversion algorithms to transform ordinal scores into interval-level scores, which provide a better way to monitor changes in parenting beliefs about infant care and more confidently use parametric statistics to evaluate the relationships between parenting beliefs and other key variables (Linacre, 2004).

The Baby Care Questionnaire

The BCQ (Winstanley & Gattis, 2013) measures parenting beliefs about infant care using two scales, structure and attunement, across different contexts such as sleeping, feeding, and soothing. The attunement scale measures belief in the value of reading and responding to infant cues and identifying infants' needs and states (e.g., hunger versus satiety, distress versus calm). The structure scale measures belief in the value of regularity and routines in infant care (e.g., time schedules for breastfeeding and sleeping). Parents are asked to rate their agreement versus disagreement using a four-point Likert scale (1 = *strongly disagree* to 4 = *strongly agree*) on each individual item of the two scales, such as "When babies cry in the night to check if someone is near, it is best to leave them" (negatively coded attunement item) and "It is important to introduce a sleeping schedule as early as possible" (positively coded structure item). Studies demonstrated good psychometric properties of the BCQ (Gattis et al., 2022; Winstanley & Gattis, 2013). The exploratory factor analysis and confirmatory factor analysis confirmed the structural validity of the two independent scales of the BCQ which were attunement and structure (Winstanley & Gattis, 2013). Studies also demonstrated that each scale achieved strong reliability and validity as measures of parenting beliefs about infant care (Gattis et al., 2022; Mascheroni et al., 2022; Winstanley & Gattis, 2013; Winstanley et al, 2014). Specifically, research indicated that both structure and attunement belief scores captured by the BCQ were related to the frequency and duration of responsive and demanding parenting behaviours (Gattis et al., 2022). In addition, attunement and structure beliefs were related to parenting experience. For attunement, pregnant women who were experienced mothers (i.e., already had at least one child) had higher scale scores compared to pregnant women expecting their first child,

while for structure, pregnant women expecting their first child had higher scale scores compared to pregnant women who were experienced mothers (Mascheroni et al., 2022).

Even though both scales of the BCQ are well-validated measures of parenting beliefs about infant care, these scales are still ordinal measures and therefore have some limitations (Tennant & Conaghan, 2007). That is, the distance between ordinal response categories of individual items, such as 1 and 2 versus 2 and 3, are not the same, meaning that an ordinal scale is unable to reflect actual change as accurately as an interval scale does (Masters, 1982; Truong et al., 2021). In addition, when requiring parametric statistics (e.g., means and standard deviations), ordinal scores cannot be used as they do not meet their arithmetic assumptions. Studies have shown that using parametric statistics with ordinal scores raises concerns about whether correct inferences can be drawn, which potentially impairs the control of Type I and Type II errors, and statistical power (Zumbo & Zimmerman, 1993; Verhulst & Neale, 2021). The use of interval-level data can minimise these concerns by improving reliability and internal validity of measurement (Jamieson, 2004; Merbitz et al., 1989). Rasch analysis (Rasch, 1960) is a method to resolve such issues and has been increasingly used to investigate psychometric properties of measures and improve their accuracy to approximate an interval-level scale (Hobart & Cano, 2009; Lundgren-Nilsson & Tennant, 2011).

Rasch Analysis

Unlike other statistical methods, such as classical test theory and generalisability theory (Cronbach et al., 1963), Rasch methodology can accurately estimate the unique contributions of individual items to the overall latent variable (e.g. structure or attunement beliefs) based on sample parameters (Fox & Jones, 1998; Rasch, 1960, 1961). A Rasch model is unidimensional and assumes that the response to a specific scale item is a function of that item's difficulty and its respondent's ability (Rasch, 1960, 1961). Rasch analysis also assumes scale invariance, which is tested by investigating Differential Item Functioning (DIF) due to personal characteristics (e.g., mother age, baby sex). DIF is useful to test whether an item of the measure works equally well across sub-groups within the population (Hagquist & Andrich, 2004).

Importantly, when the Rasch model fit is satisfactory, the ordinal scores of a psychometric measure can be converted into interval-level scores (Barber et al., 2022; Linacre, 2004; Norquist et al., 2004; Truong et al., 2023). This is the key advantage of the Rasch model over classical test theory methods because the interval-transformed data will reflect changes on a latent trait more accurately, similar to other

interval-level scales such as blood pressure or height (Hobart & Cano, 2009; Rasch, 1960; Tennant & Conaghan, 2007; Wilson, 2004; Wright & Stone, 1979). Additionally, an item-person threshold distribution plotted from the Rasch analysis is another useful tool. This graph is useful to detect possible significant ceiling or floor effects (Medvedev & Krägeloh, 2022). Thus, Rasch analysis can be considered the most advanced statistical methodology to precisely evaluate the reliability and validity of an ordinal measure, as well as enhance its precision to approximate an interval-level scale.

Our Study

Our study used Rasch methodology to evaluate the psychometric properties of the structure and attunement scales of the BCQ and enhance their precision in assessing parenting beliefs about infant care. Our study also generated conversion tables to transform ordinal structure and attunement scores to approximate interval-level scales.

Method

Participants

The optimal sample size for Rasch analysis using RUMM2030 software is between 250 and 500 cases to minimise both Type I and Type II errors (Hagell & Westergren, 2016). Hagell and Westergren (2016) suggested if the sample size is larger than 500 cases, it has been shown to inflate chi-square statistics leading to Type I error. Conversely, Type II error is likely if the samples are below 250 cases due to limited information for item calibration. These mean that sample sizes around $n = 250$ to $n = 500$ are effective in balancing the statistical interpretation of RUMM fit statistics, particularly for minimising Type I and Type II errors under the assumption of Rasch model fit.

To achieve an optimal sample size for Rasch analyses in order to investigate scale invariance between countries, we randomly selected 225 participants from each sample. Our study, with a sample size of 450 for Rasch analysis, aligns with this recommendation and is therefore well-positioned to provide reliable Rasch analysis results using the RUMM2030 software. Our sample size estimates, while based on dichotomous scales investigated by Hagell and Westergren (2016), are applicable to polytomous scales as the core principles of Rasch analysis and the Chi-square statistics used in RUMM2030 for assessing model fit are fundamentally similar across both scale types, focusing on the relationship between item difficulty and respondent ability irrespective of the number of response options.

Table 1 Demographic details of participants for Rasch sample and the UK and NZ random subsamples including their statistical comparisons using Chi-square (p-values) for categorical variables and t-tests (p-values) for continuous variables
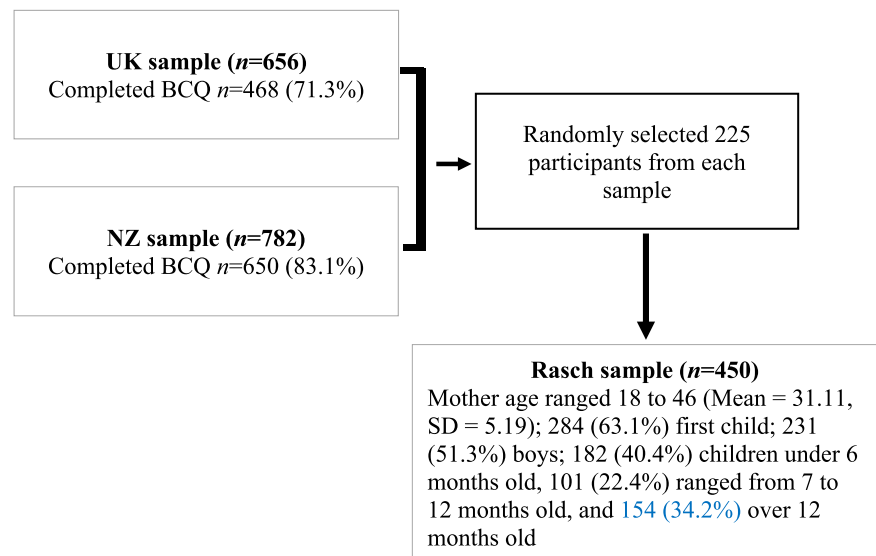
Demographic details	Rasch sample ($n = 450$)	UK random sample ($n = 225$)	NZ random sample ($n = 225$)	p^{a}
Mother age				
Mean (SD)	31.11 (5.19)	30.89 (5.45)	31.33 (4.92)	0.37
Child order				
First child	284	155	130	0.01
Not first child	164	69	95	
Child age				
Under 6 months old	182	108	74	<0.01
7 to 12 months old	101	27	74	
Over 12 months old	154	77	77	
Child gender				
Boy	231	119	112	0.51
Girl	215	104	111	

[a]Statistical comparison tests conducted between UK and NZ random subsamples only

The demographic details of the participants in Rasch sample and subsamples of each country, and the results of chi-square and statistical tests comparing differences in these demographics between subsamples are presented in Table 1. Some demographics were missing in the Rasch sample, but they were negligible. These included two mothers who did not report their child's birth order, 13 who did not report their child's age, and 4 who did not report their child's gender. Statistical comparison tests indicated that there were no significant differences for mother's age or child gender ratio between the randomly selected UK and NZ subsamples. However, in the UK subsample, there were more first-time parents and more parents of younger infants compared to the NZ subsample.

Figure 1 presents the CONSORT diagram of how participants were selected for Rasch analyses from two studies: one conducted in the United Kingdom (UK) and another in New Zealand (NZ). A CONSORT diagram was used as a standardised visual representation that can display the flow of participants through different samples. It provides a comprehensive overview of participant enrolment, allocation, follow-up, and analysis. The UK sample consisted of 656 mothers who were recruited through advertisements on parenting websites and social media including BabyCentre, Facebook, and Twitter. UK participants completed a Qualtrics survey that included brief demographics and the Baby Care Questionnaire, and they were rewarded with a

Fig. 1 CONSORT diagram for participants selected for Rasch analysis of the structure and attunement scales of the BCQ

UK sample (*n*=656)
Completed BCQ *n*=468 (71.3%)

NZ sample (*n*=782)
Completed BCQ *n*=650 (83.1%)

Randomly selected 225 participants from each sample

Rasch sample (*n*=450)
Mother age ranged 18 to 46 (Mean = 31.11, SD = 5.19); 284 (63.1%) first child; 231 (51.3%) boys; 182 (40.4%) children under 6 months old, 101 (22.4%) ranged from 7 to 12 months old, and 154 (34.2%) over 12 months old

small gift such as a baby toy or t-shirt. Out of the UK sample, the 468 mothers who completed the BCQ were 18–46 years old with a mean of 31.02 years (*SD* = 5.36). The NZ sample included 792 individuals who started a large online survey focusing on parenting around infant sleep. NZ participants were primary caregivers of an infant under two years old and were recruited via social media posts and links disseminated in news media stories. Those who were non-NZ residents (*n* = 40) or non-mothers (fathers, grandparents, and other caregivers, *n* = 12) were omitted from the sample. Out of the remaining NZ participants, 650 resident mothers who completed the BCQ were aged 19–42 with a mean age of 31.43 (*SD* = 4.95). All study procedures were reviewed and approved by authors' institution research ethics committee, which are the internationally accepted ethical standards. All participants provided informed consent.

Measure

The Baby Care Questionnaire (BCQ; Winstanley & Gattis, 2013) measures parenting beliefs about infant care using two scales: attunement and structure. Each individual item of the BCQ uses a four-point Likert scale from 1 = *strongly disagree* to 4 = *strongly agree*. The structure scale consisted of 17 items and the structure score was the mean of the structure items. The structure scale showed strong internal consistency, with McDonald's omega above 0.88 in the UK and NZ samples. The attunement scale consisted of 13 items and the attunement score was the mean of the attunement items. The attunement scale showed strong internal consistency, with McDonald's omega of above 0.83 in the UK and NZ samples.

Data Analyses

IBM SPSS v.28 was used to obtain descriptive statistics including mean and standard deviation (SD), as well as to compute McDonald's omega for the structure and attunement scales of the BCQ because McDonald's omega was not available in previous versions of IBM SPSS. RUMM2030 software package (Andrich et al., 2009) was used to conduct Rasch analyses based on the standardised criteria for the Rasch model fit as recommended elsewhere (Leung et al., 2014; Tennant & Conaghan, 2007).

Prior to Rasch analysis, the likelihood-ratio tests were conducted to examine distances between thresholds of individual items because individual items of the BCQ were polytomous. If likelihood-ratio tests showed significant differences between thresholds across individual items of both BCQ scales, the Partial Credit model (Masters, 1982) should be applied (Lundgren-Nilsson & Tennant, 2011). Otherwise, the Rating Scale model (Andrich, 1978) would be more appropriate (Tennant & Conaghan, 2007). The difference between these two polytomous-item Rasch models is that the Partial Credit model is unrestricted and assumes that each item of a scale has its own response category structure, whereas the Rating Scale model assumes that response options across all individual items of a scale have the same rating response category structure (Linacre, 2000).

Rasch analysis was then conducted iteratively until achieving the best model fit. There are several requirements for the overall Rasch model fit that generally describe parameters of an interval scale. The first requires a non-significant chi-square goodness of fit estimate of item-trait interaction (*p* > .05) as this indicates a summary of the individual item fit (Linacre, 2002). The second requires no

item misfit, which can be detected by observing fit residuals for individual items that should be in the range of ±2.50 (Lundgren-Nilsson & Tennant, 2011). The third requires no local dependency between items and can be verified by examining the residual correlations between individual items where expected values should be below 0.20 (Christensen et al., 2013). The fourth requires invariance or no DIF due to personal factors (e.g. age, sex) to ensure individual items working equally well across different groups of people (Hagquist & Andrich, 2004). Lastly is the requirement of unidimensionality, which is typically examined by the principal components analysis (PCA) of the residuals and the equating t-test (Leung et al., 2014). T-tests were conducted to compare person estimates for two sets of items, one with the highest and another with the lowest loadings, on the first principal component of residuals. Unidimensionality is evident when the amount of significant t-test comparisons does not exceed 5% or the lower bound of the binominal confidence interval calculated for significant t-tests is at or below 5% (Smith, 2002). Additionally, the reliability estimate Person Separation Index (PSI) used in Rasch analysis, although not a criterion for the Rasch model fit, reflects how well the measure discriminates between individuals with different levels of the latent trait (e.g., attunement beliefs). The interpretation of PSI is similar to Cronbach's alpha or Omega, with values of 0.70–0.79 indicating acceptable reliability and 0.80 and higher indicating good to excellent reliability (Andrich et al., 2009; Medvedev et al. 2018a, 2018b).

To ensure the consistency of item responses across various personal factors such as country, language version, or gender, we conducted Differential Item Functioning (DIF) analysis. DIF occurs when participants belonging to different sub-groups (such as individuals from different countries), despite having comparable levels of the underlying trait (like structure or attunement), exhibit distinct responses to the same item (Medvedev & Krageloh, 2022). The process of DIF testing involved the use of group-based ANOVA on standardised residuals and Bonferroni-adjusted t-tests for specific subgroup comparisons, as suggested by Andrich & Marais (2019). Furthermore, we visually inspected individual item plots, namely the item characteristic curves (ICCs). In cases where significant DIF was detected for any item or testlet (as detailed in the following paragraph), a thorough visual analysis of the plot was undertaken to identify whether the DIF was uniform (indicating negligible group overlap across all trait levels) or nonuniform (indicating notable overlap).

In this study, we considered combining individual items into testlets to reduce measurement error due to local dependency, DIF and item misfit and to improve the overall Rasch model fit; this is a well-established methodology (Lundgren-Nilsson et al., 2013; Medvedev et al., 2018a,

2018b; Merkin et al., 2020; Truong et al., 2023; Truong, Numbers, et al., 2023). Testlets were created and combined based on local dependency reflected by residual correlations exceeding 0.20 in magnitude because local dependency produces spurious correlations and artificially affects model fit and dimensionality. We have considered the magnitude of residual correlations on the first place when creating testlets. The effect of combining individual items into testlets to reduce undesired error variances is similar to the signal-averaging technique used to study event-related potentials using electroencephalogram (EEG). In this technique, the test is repeated several times until error variances nullify each other, allowing researchers to identify the relevant neural response (Luck, 2014). Similarly, instead of replicating the test, combining multiple items that measure the same construct can eliminate unwanted variances that are not related to the construct being measured (Truong et al., 2023). Therefore, if the best fit to the Rasch model is achieved after creating testlets, it indicates that measurement errors were successfully decreased using this approach.

When the data fit the Rasch model, the person-item threshold distribution was examined to see how well item thresholds of each BCQ scale covered the levels of parenting beliefs about infant care in the sample. Finally, conversion tables for the structure and attunement scales of the BCQ were generated to convert ordinal raw scores into interval-level data.

Results

The Unrestricted Rasch Partial Credit model (Masters, 1982) was used for the data of this study because likelihood-ratio tests showed significant differences between thresholds across individual items of both BCQ scales (p's < 0.001).

Structure Scale

Table 2 displays the overall model fit estimates of the initial and final Rasch analyses of the structure scale of the BCQ. As can be seen, the initial analysis indicated reasonable reliability (PSI = 0.67) for the structure scale but the overall fit to the Rasch model was unacceptable due to significant chi-square, $\chi^2(153) = 626.01$, $p < 0.001$, and lack of evidence for unidimensionality (13.1% of significant t-tests). Table 3 displays individual item statistics including location, item-fit residual, and chi-square values for the initial analysis of the structure scale of the BCQ. There are two items with significant misfit to the model: item S5 and S9.

The residual correlation matrix revealed local dependency for several groups of items as indicated by their

Table 2 Summary of fit statistics for the model fit Rasch analyses of the structure of the BCQ

Analysis	Person mean		Goodness of fit		PSI	Significant t-tests (Unidimensionality)		A§
	Value	SD	χ2 (df)	p		%	Lower bound	
Structure								
Initial	0.00	1.59	626.02 (153)	<0.001	0.67	15.1	13.1 (NO)	–
Final	0.08	0.90	5.66 (12)	0.93	0.78	7.1	5.0 (YES)	1.16
Attunement								
Initial	0.71	1.12	602.95 (117)	<0.001	0.85	18.4	13.1 (NO)	–
Final	0.86	1.89	4.89 (8)	0.77	0.92	6.9	4.9 (YES)	1.08

§A indicates the retained common variance in the estimate

Table 3 Rasch model fit statistics of item-fit residuals for the initial analysis of the individual items in the structure and attunement scales of the BCQ

Scale/item	Item location	Item-fit residual	Chi-square
Structure			
S1. Babies can have a good night's sleep regardless of scheduling	0.62	0.78	7.42
S2. Strict sleeping routines prevent parent(s) from enjoying their child	−0.47	−1.47	17.86
S3. Sleeping schedules make babies unhappy	−1.01	−0.67	19.00
S4. It is important to introduce a sleeping schedule as early as possible	0.16	−2.46	26.25
S5. Babies benefit from a quiet room to sleep	0.43	3.15*	61.42
S6. Babies benefit from a fixed napping/sleeping schedule	0.16	−0.74	7.80
S7. Implementing feeding/eating schedules leads to a calm and content baby	0.23	−1.91	23.15
S8. Feeding/eating routines are difficult to follow	0.25	−0.81	11.93
S9. One danger of feeding/eating schedules is that babies might not get enough to eat	−0.16	8.76*	270.89
S10. Following feeding/eating routines prevents parent(s) from enjoying parenthood to the full	−0.42	−1.03	9.80
S11. It is important to introduce a feeding/eating schedule as early as possible	0.13	−1.14	53.16
S12. Babies will not follow feeding/eating schedules	−0.41	−0.11	17.20
S13. Babies with regular schedules spend less time crying	0.16	−1.75	17.99
S14. Babies cry no matter what their routines	0.72	1.69	38.68
S15. Routines lead to more crying	−0.91	−1.00	15.04
S16. Having a set routine helps an upset baby calm down	0.01	−0.13	12.07
S17. Babies with regular schedules cry just as much as babies without regular schedules	0.52	0.46	16.37
Attunement			
A1. Some days, babies need more or less sleep than other days	0.67	−3.50*	50.39
A2. Babies benefit from physical contact with parent(s) when they wake during the night	0.50	−5.23*	42.75
A3. When babies cry in the night to check if someone is near, it is best to leave them	−0.70	2.30	87.68
A4. Parent(s) should find a pattern of feeding/eating that suits the baby	0.50	−3.05*	52.27
A5. Baby-led feeding leads to behavioural and sleep problems	−0.77	2.35	69.36
A6. Offering milk/food to a baby is a good way to test whether she/he is hungry	0.67	−1.15	19.04
A7. Babies will eat whenever milk/food is offered even if they are not hungry	−0.74	2.79	61.25
A8. Parent(s) should delay responding to a crying baby	−1.19	1.21	36.92
A9. It is a good idea to have a set time you leave a baby to calm herself/himself down, and increase this amount of time each week	−0.44	2.15	66.95
A10. Physical contact such as stroking or rocking helps a baby to be calm	0.69	−4.12*	32.48
A11. Holding babies frequently during the day makes them more demanding	−0.48	0.68	32.53
A12. Responding quickly to a crying baby leads to less crying in the long run	0.58	−2.58*	22.36
A13. Leaving a baby to cry can cause emotional insecurity	0.71	−3.89*	28.98

*Significant misfit to the Rasch model

Table 4 Rasch model fit statistics of item-fit residuals for the final analysis of the testlets in the structure and attunement scales of the BCQ

Scale/ item	Item location	Item-fit residual	Chi-square
Structure			
Testlet 1 (items S3, S4, S5, S9, S10 & S11)	−0.08	0.66	2.99
Testlet 2 (items S2, S8, & S16)	0.00	1.29	1.52
Testlet 3 (items S1, S6, S7, S12, S13, S14, S15, & S17)	0.08	−2.31	1.14
Attunement			
Testlet 1 (items A1, A4, A6, A7, A8, A9, A12, & A13)	−0.17	−1.94	0.83
Testlet 2 (items A2, A3, A5, A10, & A11)	0.17	1.24	4.06
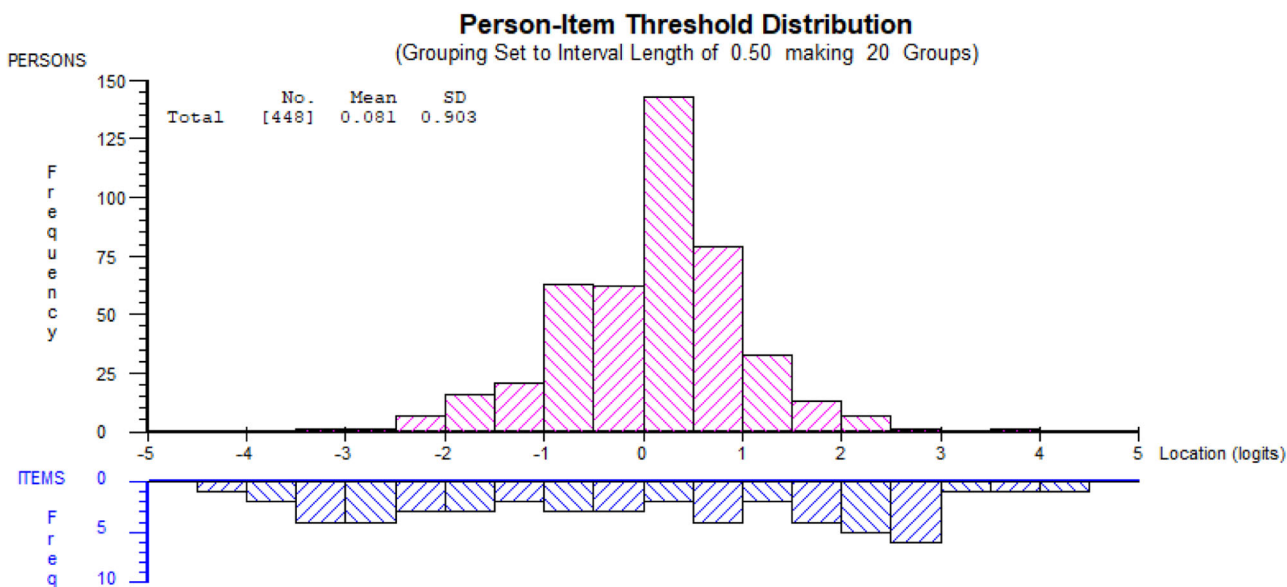
Fig. 2 Person-item threshold distribution of the model fit analysis of the structure scale of the BCQ

residual correlations above 0.20. In addition, many structure items had DIF issues and especially DIF by country in several structure items (i.e., items S1, S4, S5, S9, S11, S13, S14, and S16). This means that such structure items did not work equally well in both the UK and NZ subsamples. Local dependency between items affects the model fit and may impact on DIF or scale invariance (Christensen et al., 2013). Several studies have demonstrated the impact of local dependency on Rasch model fit and DIF. For example, Adams and Wu (2002) showed that local dependency can lead to inflated estimates of item and person parameters, and can result in spurious DIF. Similarly, Linacre (2006) demonstrated that local dependency can lead to biased item difficulty estimates and incorrect classification of persons. To resolve local dependency issues, three testlets were created: Testlet 1 (items S3, S4, S5, S9, S10 & S11); Testlet 2 (items S2, S8, & S16) and Testlet 3 (items S1, S6, S7, S12, S13, S14, S15, & S17). This analysis resulted in acceptable unidimensionality (5.0% of significant t-tests) with retained high common variance in the estimate of $A = 1.16$, supporting unidimensionality of the obtained bifactor solution. This analysis also increased reliability

(PSI = 0.78), as well as no local dependency, no item misfit (see Table 4, Structure) and no DIF by personal factors. Therefore, this final analysis achieved the best Rasch model fit for the structure scale of the BCQ.

The person-item threshold distribution for the structure scale from the analysis of the best model fit (Final analysis) is presented in Fig. 2. This shows that the structure scale's thresholds satisfactorily cover the levels of parenting beliefs about infant care in the sample and there are no significant ceiling or floor effects. Table 5 displays the ordinal-to-interval conversion table for the structure scale of the BCQ based on person estimates of the final fit to the Rasch model.

Attunement Scale

Table 2 also shows the overall model fit estimates of the initial analysis, and final Rasch analyses of the attunement scale of the BCQ. As can be seen, initial analysis indicated strong reliability of PSI = 0.85 for the attunement scale but the overall fit to the Rasch model was unacceptable due to significant chi-square, $\chi^2(117) = 602.95$, $p < 0.001$, and a

Table 5 Converting ordinal scores into interval-level scores for the structure scale of the BCQ

Raw scores	Logits	Interval scores	Raw scores	Logits	Interval scores
1.00	−5.21	1.00	2.53	0.20	2.56
1.06	−4.56	1.19	2.59	0.39	2.62
1.12	−4.11	1.32	2.65	0.57	2.67
1.18	−3.79	1.41	2.71	0.75	2.72
1.24	−3.55	1.48	2.76	0.93	2.77
1.29	−3.35	1.54	2.82	1.11	2.83
1.35	−3.17	1.59	2.88	1.28	2.88
1.41	−3.01	1.64	2.94	1.45	2.92
1.47	−2.86	1.68	3.00	1.62	2.97
1.53	−2.71	1.72	3.06	1.78	3.02
1.59	−2.57	1.76	3.12	1.93	3.06
1.65	−2.42	1.81	3.18	2.07	3.10
1.71	−2.27	1.85	3.24	2.19	3.14
1.76	−2.11	1.90	3.29	2.30	3.17
1.82	−1.95	1.94	3.35	2.40	3.20
1.88	−1.79	1.99	3.41	2.50	3.23
1.94	−1.62	2.04	3.47	2.59	3.25
2.00	−1.45	2.09	3.53	2.68	3.28
2.06	−1.28	2.14	3.59	2.79	3.31
2.12	−1.10	2.19	3.65	2.91	3.34
2.18	−0.92	2.24	3.71	3.07	3.39
2.24	−0.73	2.29	3.76	3.28	3.45
2.29	−0.55	2.35	3.82	3.58	3.54
2.35	−0.36	2.40	3.88	3.97	3.65
2.41	−0.17	2.45	3.94	4.49	3.80
2.47	0.01	2.51	4.00	5.18	4.00

lack of unidimensionality (13.1% of *t*-tests were significant). Table 3 also displays statistics of item location, item-fit residual, and chi-square values for the initial analysis of the individual items in the attunement scale. There were several items with significant misfit to the model: item A1, A2, A4, A7, A10, A12, and A13. In addition, most attunement items had DIF issues by country factor, except the attunement item A12. Similarly, to the structure scale analysis, we applied the testlet approach as it can resolve both item misfit and DIF.

Two testlets of the attunement scale were thus created by considering residual correlations between items: Testlet 1 (items A1, A4, A6, A7, A8, A9, A12, & A13), and Testlet 2 (A2, A3, A5, A10, & A11). This resulted in the best overall model fit (Table 2, Attunement, Final analysis) with no local dependency, no item misfit (see Table 4, Attunement), no DIF amongst testlets, acceptable unidimensionality (4.9% of significant *t*-tests) and retained large common variance in the estimate of $A = 1.08$, supporting unidimensionality of the current bifactor solution with testlets,

as well as improved reliability (PSI = 0.92) compared to the initial analysis.

The person-item threshold distribution for the attunement scale from the analysis of the best model fit (Final analysis) is presented in Fig. 3. This shows that the attunement scale's thresholds satisfactorily cover the levels of parenting beliefs about infant care in the sample and there are no significant ceiling or floor effects. Table 6 displays the ordinal-to-interval conversion table for the attunement scale of the BCQ based on person estimates of the final fit to the Rasch model.

Rasch Interval-Transformed Scores

Table 7 includes the range scores, means and standard errors of the BCQ ordinal raw scores and interval-transformed Rasch scores for the UK and NZ original samples. As can be seen, all standard errors of ordinal raw scores of both structure and attunement scales were consistently higher compared to Rasch transformed scores in both original samples. In addition, there were significant differences on attunement ordinal raw scores and interval-transformed Rasch scores between the original UK and NZ samples (p's < 0.001). However, attunement interval-transformed scores (Cohen's $d = 2.69$) showed lower effect size compared to ordinal raw scores (Cohen's $d = 3.44$). These findings together illustrate that Rasch transformed scores indicate a reduction of the magnitude of the difference and thus this will derive from the smaller standard error.

Discussion

This study utilised Rasch methodology to evaluate and improve the psychometric properties of the structure and attunement scales of the Baby Care Questionnaire (BCQ), a well-established ordinal measure of parenting beliefs about infant care. The use of Rasch methodology allowed for the identification of locally dependent items that were combined into testlets to enhance the precision of both scales of the BCQ. We demonstrated that these two BCQ scales, reorganised into testlets, met expectations of the Rasch model, which then allowed us to enhance the accuracy of each scale comparable to interval-level measurement using the ordinal-to-interval conversion algorithms presented in Tables 4 and 5 for the structure and attunement scales respectively. The resulting conversion algorithms can be used to transform the ordinal scores obtained from the BCQ into interval data, which can increase the precision of the measure without changing its original response format. This can be particularly useful in longitudinal studies that track changes in parenting beliefs over time, as well as in cross-
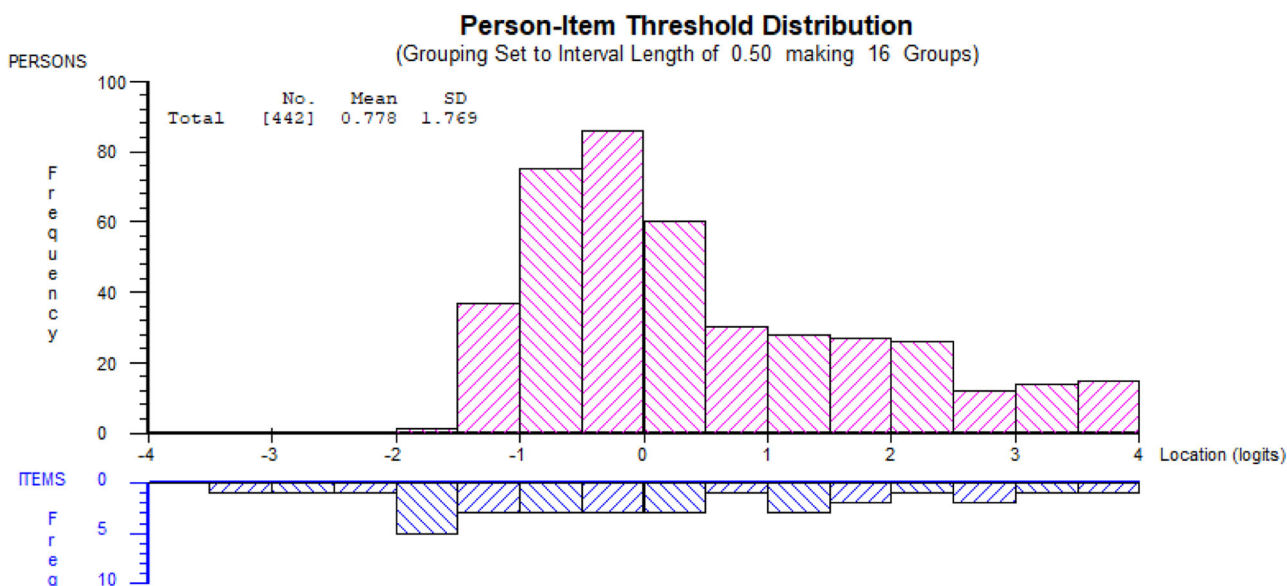
Fig. 3 Person-item threshold distribution of the model fit analysis of the attunement scale of the BCQ

Table 6 Converting ordinal scores into interval-level scores for the attunement scale of the BCQ

Raw scores	Logits	Interval scores	Raw scores	Logits	Interval scores
1.00	−7.61	1.00	2.54	−0.19	2.54
1.08	−6.44	1.24	2.62	0.01	2.58
1.15	−5.39	1.46	2.69	0.21	2.62
1.23	−4.46	1.65	2.77	0.41	2.66
1.31	−3.64	1.82	2.85	0.62	2.70
1.38	−2.94	1.97	2.92	0.84	2.75
1.46	−2.24	2.11	3.00	1.07	2.80
1.54	−1.95	2.17	3.08	1.32	2.85
1.62	−1.82	2.20	3.15	1.58	2.90
1.69	−1.72	2.22	3.23	1.86	2.96
1.77	−1.64	2.24	3.31	2.15	3.02
1.85	−1.57	2.25	3.38	2.47	3.09
1.92	−1.50	2.27	3.46	2.80	3.15
2.00	−1.42	2.28	3.54	3.14	3.23
2.08	−1.33	2.30	3.62	3.52	3.30
2.15	−1.21	2.32	3.69	3.93	3.39
2.23	−1.04	2.36	3.77	4.42	3.49
2.31	−0.82	2.41	3.85	5.03	3.62
2.38	−0.60	2.45	3.92	5.83	3.78
2.46	−0.39	2.49	4.00	6.88	4.00

cultural research that compare parenting beliefs across different populations.

These ordinal-to-interval conversions are important because individual items of each scale contribute differently to the total score of each scale, which should be accounted

for (Stucki et al., 1996). Sandham et al. (2019) used a metaphor of squeezing Vitamin C from different fruits to explain how the Rasch model works. Each fruit represents an item on the scale, and each fruit contains differing levels of Vitamin C, which represents the latent trait being measured (e.g., parenting beliefs about infant care). Just as different fruits contribute different amounts of Vitamin C to a smoothie, different items on a scale contribute different amounts of the latent trait to the overall score. Rasch analysis allows for the measurement of the latent construct to the extent that it is contained within each item while filtering out other constructs, manifesting as fit residuals in the model. This increases the precision of measurement and allows for comparisons of accuracy between original ordinal-level scores and their Rasch-converted interval equivalents (Norquist et al., 2004).

Literature reviews on Rasch methodology concluded that using Rasch interval-transformed scores can reduce measurement error associated with ordinal scores (Truong et al., 2023). Therefore, interval-transformed scores may reflect the levels of parenting beliefs about infant care more accurately, as well as avoiding the violation of arithmetic assumptions when conducting parametric statistical tests (Leung, 2011). In addition, achieving adequate reliability (PSI = 0.78) for the structure scale and excellent reliability (PSI = 0.92) for the attunement scale adds empirical evidence to support the robust reliability of these BCQ scales in measuring parenting beliefs about infant care. Interval-transformed data can also be appropriate for conducting statistical comparisons with other interval data, such as electrophysiological or neuroimaging data, contributing to improvement of reliability and validity of research results.

Table 7 Statistical comparisons between the original UK ($n = 468$) and NZ ($n = 650$) samples on ordinal raw scores and interval-transformed scores of the BCQ scales

Scores	Sample	Structure						Attunement					
		Minimum	Maximum	Mean (SD)	SE	p	d^a	Minimum	Maximum	Mean (SD)	SE	p	d^a
Ordinal													
	UK	1.06	3.76	2.46 (0.46)	0.022	0.59	0.03	1.08	2.77	1.79 (0.39)	0.019	<0.001	3.44
	NZ	1.41	3.82	2.48 (0.38)	0.016			2.00	3.92	3.20 (0.41)	0.017		
Rasch													
	UK	1.19	3.45	2.49 (0.38)	0.019	0.74	0.02	1.24	2.66	2.16 (0.27)	0.013	<0.001	2.69
	NZ	1.64	3.54	2.50 (0.32)	0.014			2.28	3.78	2.99 (0.34)	0.014		

[a]Cohen's d

Therefore, the assumptions of parametric statistical tests can be met using the interval BCQ data (Leung, 2011). Transforming the ordinal scores of both structure and attunement scales of the BCQ into interval scores does not require users to be experts in statistics because we have developed conversion algorithms based on Rasch model estimates. To interpret ordinal-level scores for each scale in our tables (i.e., structure scale in Table 5, attunement scale in Table 6), researchers can use corresponding interval scale scores found on the right-hand side of each table (Tables 5 and 6).

The findings from this study have significant implications for research in the field of child development and developmental psychology. Parenting beliefs about infant care can potentially influence parental behaviours, which can in turn impact child development. Therefore, having a reliable and valid measure of parenting beliefs is crucial for understanding the complex interactions between parenting beliefs and behaviours and child outcomes. After establishing scale invariance, we have confirmed that NZ parents score significantly higher on attunement beliefs compared to the UK sample, while there were no significant differences on structure beliefs between countries. This finding may be due to demographic differences in the samples as there were significantly more first-time mothers in the UK sample. Previous research indicates that experienced mothers report stronger beliefs in attunement compared to first-time parents (Mascheroni et al., 2022).

The main strength of this study was the application of robust Rasch methodology to evaluate and enhance the psychometric properties of the structure and attunement scales of the BCQ using a randomly selected subsample from two countries (UK and NZ), and then the sample size ($n = 450$) satisfied the optimal sample size for Rasch analyses to minimise Type I and Type II errors (Hagell & Westergren, 2016). In addition, the initial analyses detected several items has DIF issues by personal factors (i.e., mother age, infant age, and infant gender), and especially by countries. However, this study found that the modified structure and attunement scales of the BCQ are invariant,

working equally well across personal factors (i.e., mother age, infant age, and infant gender), as well as across NZ and UK mothers, which had not been investigated in previous studies. Scale invariance (no DIF) refers to the property of a measurement tool that ensures that the relationships between the items on the scale are consistent across different groups or populations. In other words, it ensures that the same construct is being measured in the same way across different groups, regardless of their culture, language, or background. When a measurement tool is invariant across groups, it means that the scores obtained from different groups can be compared and interpreted meaningfully. Moreover, in our analysis, we observed that interval-transformed Rasch scores showed a smaller effect size compared to ordinal raw scores. This observation reflects the impact of Rasch transformation on the standard deviation component of effect size calculation, rather than indicating a direct reduction in measurement error. Rasch transformation, by converting ordinal data to an interval scale, affects the distribution and variability of scores, which in turn influences the effect size. This transformation enhances the interpretability and linearity of the scores but does not inherently imply a reduction in measurement error.

We noted that there were significant differences on both ordinal raw scores and interval-transformed Rasch attunement scores between the original UK and NZ samples with interval-transformed attunement scores showed smaller effect size compared to its ordinal raw scores. However, there were no significant differences on both ordinal raw scores and interval-transformed Rasch structure scores between the original UK and NZ samples with interval-transformed structure scores showed slightly lower effect size compared to its ordinal raw scores. These may indicate that structure is more similar across cultures (UK vs NZ) compared to attunement, where cultural differences may affect the accuracy of the ordinal assessment scale. However, invariance was established by Rasch modifications for both scales, which is reflected by differences in effect sizes

between ordinal and Rasch scores for attunement but not for structure.

When we compare parenting beliefs about structure and attunement between different countries, we need to ensure that the measurement tool (e.g., BCQ) is scale-invariant, so that we can trust that any differences between countries are real and not due to measurement bias. Even if the mean scores between countries are significantly different, we cannot conclude that there is a real difference unless we can demonstrate that the measurement tool is invariant across countries, which we achieved in this study. For example, suppose we are comparing the parenting beliefs of parents from two different countries (UK and NZ) using a scale that has been shown to be invariant across these cultures. If the mean score for parents from the UK is significantly higher than that of parents from NZ, we can interpret this difference as reflecting a real difference in parenting beliefs between the participants from the two samples. As acknowledged earlier, there were more experienced mothers in the NZ sample, potentially accounting for the higher attunement scores in the NZ sample. Therefore, the country difference observed in this study may not reflect a real difference between NZ and UK mothers at large. However, if the scale is not invariant across cultures and DIF occurs, then any differences in mean scores may be due to measurement bias rather than actual differences in parenting beliefs between samples or countries. Scale invariance is essential when comparing parenting beliefs or any other constructs across different cultures or populations. Only when we can demonstrate that the measurement tool is invariant across cultures can we reliably compare mean scores between groups and draw meaningful conclusions about differences in parenting beliefs.

This study is not without limitations. Participants in this study came from convenience samples in two countries and may not be representative of mothers within those two countries. In addition, there are other dimensions that might be relevant to parenting beliefs, besides parent and infant age and infant gender. Therefore, replications of this study should be conducted in samples from other English-speaking countries (and possibly involving translations to other languages), and should consider other potentially relevant personal factors, such as differences among cultures, parenting experience, levels of educational attainment, partnership status of parents, and gender of parents.

In conclusion, the findings of this study demonstrated the reliability and internal validity of both scales of the BCQ in measuring parenting beliefs about infant care. Our minor modifications implemented in the scoring of both scales of the BCQ satisfied expectations of the unidimensional Rasch model and scale invariance across different country samples, infant and parent age and infant gender, as well as resolved local dependency issues. This allowed us to enhance the accuracy of each scale to approximate interval-level scores using the ordinal-to-interval conversion tables. Researchers can use the conversion tables published here to enhance precision of scores on the structure and attunement scales of the BCQ without changing their original questionnaire format. Overall, the findings from this study contribute to the enhancement of precision in measuring parenting beliefs about infant care, which is important for understanding the complex interactions between parents and their infants and for developing interventions to promote positive parenting practices and support healthy child development.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and data analyses were performed by Q.C.T, O.N.M, M.G., and C.C.B. The first draft of the manuscript was completed by Q.C.T., and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

Ethical approval All study procedures were reviewed and approved by the research ethics committees of Cardiff University and University of Waikato and adhere to the tenets of the Declaration of Helsinki including informed consent. Study data are available at https://osf.io/7c9r5?view_only=522b3f90add441549312489468b6d0e6.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Adams, R. J., & Wu, M. L. (2002). *Programme for international student achievement: technical report*. Paris: Organisation for Economic Cooperation and Development.

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

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: measuring in the educational, social and health sciences*. Springer Nature Singapore Pte Ltd.

Andrich, D., Sheridan, B., & Luo, G. (2009). *RUMM 2030 (Beta Version for Windows) In Perth*. Western Australia: RUMM Laboratory Pty Ltd.

Barber, C. C., Middlemiss, W., & Medvedev, O. N. (2022). Applying Rasch methodology to examine and enhance precision of the Edinburgh Postnatal Depression Scale. *Journal of Affective Disorders*, *308*, 391–397. https://doi.org/10.1016/j.jad.2022.04.009.

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch models in health*. Wiley Online Library.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, *45*(1), 30–45.

Galbally, M., Watson, S. J., Teti, D., & Lewis, A. J. (2018). Perinatal maternal depression, antidepressant use and infant sleep outcomes: exploring cross-lagged associations in a pregnancy cohort study. *Journal of Affective Disorders*, *238*, 218–225. https://doi.org/10.1016/j.jad.2018.05.025.

Gattis, M., Winstanley, A., & Bristow, F. (2022). Parenting beliefs about attunement and structure are related to observed parenting behaviours. *Cogent Psycholology*, *9*(1), Article 2082675. https://doi.org/10.1080/23311908.2022.2082675.

Hagell, P., & Westergren, A. (2016). Sample size and statistical conclusions from tests of fit to the Rasch model according to the Rasch Unidimensional Measurement Model (RUMM) Program in Health Outcome Measurement. *Journal of Applied Measurement*, *17*(4), 416–431.

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences*, *36*(4), 955–968. https://doi.org/10.1016/S0191-8869(03)00164-8.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment*, *13*(12), iii–177. https://doi.org/10.3310/hta13120.

Hughes, S. O., Cross, M. B., Hennessy, E., Tovar, A., Economos, C. D., & Power, T. G. (2012). Caregiver's feeding styles questionnaire. Establishing cutoff points. *Appetite*, *58*(1), 393–395. https://doi.org/10.1016/j.appet.2011.11.011.

Jamieson, S. (2004). Likert scales: how to (ab) use them? *Medical Education*, *38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x.

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research*, *37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697.

Leung, Y.-Y., Png, M.-E., Conaghan, P., & Tennant, A. (2014). A systematic literature review on the application of Rasch analysis in musculoskeletal disease—a special interest group report of OMERACT 11. *Journal of Rheumatology*, *41*(1), 159–164. https://doi.org/10.3899/jrheum.130814.

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

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

Linacre, J. M. (2004). Estimation methods for Rasch measures. In E. V. Smith Jr, & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 25-48). JAM Press.

Linacre, J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement*, *7*(1), 129–139.

Lundgren-Nilsson, A., Jonsdottir, I. H., Ahlborg, G., & Tennant, A. (2013). Construct validity of the psychological general well being index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: a Rasch analysis. *Health and Quality of Life Outcomes*, *11*(1), Article 2. https://doi.org/10.1186/1477-7525-11-2.

Lundgren-Nilsson, A., & Tennant, A. (2011). Past and present issues in Rasch analysis: the functional independence measure (FIM™) revisited. *Journal of Rehabilitation Medicine*, *43*(10), 884–891. https://doi.org/10.2340/16501977-0871.

Mascheroni, E., Grassi, M., Bonanomi, A., Sperotto, R., Deeg, S., Hung, S., Xia, R., Ionio, C., Au, T. K. F., & Gattis, M. (2022). The role of experience in parenting beliefs of British and Italian women during pregnancy. *Infant Mental Health Journal*, *43*(6), 835–848. https://doi.org/10.1002/imhj.22014.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272.

Medvedev, O. N., & Krägeloh, C. U. (2022). Rasch measurement model. In Medvedev, O.N., Krägeloh, C.U., Siegert, R.J., Singh, N.N. (Eds), *Handbook of assessment in mindfulness research* (pp. 1–18). Springer.

Medvedev, O. N., Titkova, E. A., Siegert, R. J., Hwang, Y. S., & Krageloh, C. U. (2018a). Evaluating short versions of the five facet mindfulness questionnaire using Rasch analysis. *Mindfulness*, *9*(5), 1411–1422. https://doi.org/10.1007/s12671-017-0881-0.

Medvedev, O. N., Turner-Stokes, L., Ashford, S., & Siegert, R. J. (2018b). Rasch analysis of the UK Functional Assessment Measure in patients with complex disability after stroke. *Journal of Rehabilitation Medicine*, *50*(5), 420–428. https://doi.org/10.2340/16501977-2324.

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

Merkin, A. G., Medvedev, O. N., Sachdev, P. S., Tippett, L., Krishnamurthi, R., Mahon, S., Kasabov, N., Parmar, P., Crawford, J., & Doborjeh, Z. G. (2020). New avenue for the geriatric depression scale: Rasch transformation enhances reliability of assessment. *Journal of Affective Disorders*, *264*, 7–14. https://doi.org/10.1016/j.jad.2019.11.100.

Norquist, J. M., Fitzpatrick, R., Dawson, J., & Jenkinson, C. (2004). Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Medical Care*, 125–136. https://doi.org/10.1097/01.mlr.0000103530.13056.88.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In Neyman, J. (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. University of California.

Sandham, M. H., Medvedev, O. N., Hedgecock, E., Higginson, I. J., & Siegert, R. J. (2019). A Rasch analysis of the integrated palliative care outcome scale. *Journal of Pain and Symptom Management*, *57*(2), 290–296. https://doi.org/10.1016/j.jpainsymman.2018.11.019.

Smetana, J. G. (2017). Current research on parenting styles, dimensions, and beliefs. *Curr Opin Psychol*, *15*, 19–25. https://doi.org/10.1016/j.copsyc.2017.02.012.

Smith, Jr, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Stucki, G., Daltroy, L., Katz, J. N., Johannesson, M., & Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of Clinical Epidemiology*, *49*(7), 711–717. https://doi.org/10.1016/0895-4356(96)00016-9.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358–1362. https://doi.org/10.1002/art.23108.

Tikotzky, L., & Sadeh, A. (2009). Maternal sleep-related cognitions and infant sleep: a longitudinal study from pregnancy through the 1st year. *Child Dev*, *80*(3), 860–874. https://doi.org/10.1111/j.1467-8624.2009.01302.x.

Truong, Q. C., Choo, C., Numbers, K., Bentvelzen, A., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2023). Enhancing precision of the telephone interview for cognitive status-modified (TICS-M) using the Rasch model. *Psychological Assessment*. https://doi.org/10.1037/pas0001233.

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2021). Enhancing precision of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16) using Rasch methodology. *International Psychogeriatrics*, 1–11. https://doi.org/10.1017/S1041610221002568.

Truong, Q. C., Numbers, K., Choo, C. C., Bentvelzen, A. C., Catts, V. S., Cervin, M., Jorm, A. F., Kochan, N. A., Brodaty, H., Sachdev, P., & Medvedev, O. N. (2023). Establishing conversion of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly scores into interval-level data across multiple samples using Rasch methodology. *Psychogeriatrics*. https://doi.org/10.1111/psyg.12946.

Verhulst, B., & Neale, M. C. (2021). Best practices for binary and ordinal data analyses. *Behavior Genetics*, *51*(3), 204–214. https://doi.org/10.1007/s10519-020-10031-x.

Wilson, M. (2004). *Constructing measures: an item response modeling approach*. Routledge.

Winstanley, A., & Gattis, M. (2013). The Baby Care Questionnaire: a measure of parenting principles and practices during infancy. *Infant Behavior & Development*, *36*(4), 762–775. https://doi.org/10.1016/j.infbeh.2013.08.004.

Winstanley, A., Sperotto, R. G., Putnick, D. L., Cherian, S., Bornstein, M. H., & Gattis, M. (2014). Consistency of maternal cognitions and principles across the first five months following preterm and term deliveries. *Infant Behavior & Development*, *37*, 760–771. https://doi.org/10.1016/j.infbeh.2014.09.005.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, *34*(4), 390–400. https://doi.org/10.1037/h0078865.