Quality of Life Research (2022) 31:1–9
https://doi.org/10.1007/s11136-021-03022-w

SPECIAL SECTION: NON-PARAMETRIC IRT

Advances in nonparametric item response theory for scale construction in quality-of-life research

Klaas Sijtsma[1] · **L. Andries van der Ark**[2]

Accepted: 13 October 2021 / Published online: 9 November 2021

Abstract

We introduce the special section on nonparametric item response theory (IRT) in *Quality of Life Research*. Starting from the well-known Rasch model, we provide a brief overview of nonparametric IRT models and discuss the assumptions, the properties, and the investigation of goodness of fit. We provide references to more detailed texts to help readers getting acquainted with nonparametric IRT models. In addition, we show how the rather diverse papers in the special section fit into the nonparametric IRT framework. Finally, we illustrate the application of nonparametric IRT models using data from a questionnaire measuring activity limitations in walking. The real-data example shows the quality of the scale and its constituent items with respect to dimensionality, local independence, monotonicity, and invariant item ordering.

Keywords Goodness of fit · Measurement of health-related attributes · Nonparametric item response theory · Rasch model

Introduction

This special section of *Quality of Life Research* is devoted to nonparametric item response theory (IRT) models [1]. In this introduction, we review nonparametric IRT models, provide references to more detailed texts, and show how the diverse set of papers in the special section fits into the nonparametric IRT framework. Nonparametric IRT models are generalizations of a large class of parametric IRT models including the Rasch model, the 2-parameter and 3-parameter logistic IRT models for binary item scores, and the partial credit model and the graded response model for polytomous item scores. Van der Linden [2] introduces these parametric IRT models extensively. Sijtsma and Van der Ark [3, Chap. 4] discussed how nonparametric and parametric IRT models are related in one large family of which the most general members are nonparametric IRT models. Parametric IRT models are special cases of nonparametric IRT models. Their generality renders nonparametric IRT models more flexible than most parametric IRT models.

IRT models are used for establishing whether a set of items intended to measure a particular attribute together constitute a scale for measurement. Examples of attributes are pain experienced by patients suffering from burn wounds [4], health-related quality-of-life aspects, such as physical functioning, general health perceptions, vitality, and social functioning [5], and adherence to medication and lifestyle for patients with hypertension [6]. Roorda et al. [7] used IRT models for scaling a set of items measuring activity limitations in rising and sitting down in patients with lower-extremity disorders living at home, and Sijtsma et al. [8] used nonparametric IRT models to analyze the World Health Organization Quality-of-Life scale (WHOQOL-Bref). In the special section, Feng et al. [9] applied nonparametric IRT models to the EQ-5D, a widely used generic measure of health, and based on the scaling results reinterpreted the EQ-5D scales. The number of articles using IRT models and other scaling techniques for constructing scales and assessing measurement quality *Quality of Life Research* published over the years is very large. It shows the paramount importance of well-founded measurement.

✉ L. Andries van der Ark
L.A.vanderArk@uva.nl

[1] Tilburg School of Social and Behavioural Sciences, Tilburg University, P.O Box 90153, 5000 LE Tilburg, The Netherlands

[2] Research Institute of Child Development and Education, University of Amsterdam, P. O. Box 15776, 1001 NG Amsterdam, The Netherlands

Features or nonparametric and parametric IRT modeling

The main difference between nonparametric and parametric IRT models is that the nonparametric models rest on assumptions about people responding to items in a test or a questionnaire that are more liberal than the assumptions parametric models make. For example, nonparametric IRT models assume that the relation between the probability of a patient giving a positive response to an item indicating ease of climbing the stairs and the underlying attribute of physical functioning is monotone—the better physical functioning, the more ease climbing the stairs—and parametric IRT models assume the relation not only is monotone but also logistic. This extra condition renders the relation more restrictive and the fit of the IRT model to the data more problematic. For example, the nonparametric model of monotone homogeneity [10; 1, Chap. 3] assumes monotone item response functions (IRFs), which can have any shape and intersect mutually (Fig. 1, all curves), and the logistic IRFs of the Rasch model all have the typical S shape while running parallel (Fig. 1, dashed curves). The models are equal in that sets of items comprising a test or questionnaire measure one attribute such as physical functioning and represented mathematically by one latent variable (typically denoted by θ; Fig. 1). In addition, the models assume that there are no other attributes or covariates active affecting the covariances between item scores, so that inter-item covariances vanish when conditioning on the latent variable. These assumptions are unidimensionality and local independence, respectively. Parametric IRT models generalize to multiple latent variables thus allowing various attributes simultaneously to affect response probabilities (Fig. 2). Nonparametric IRT focuses on search algorithms to identify separate item clusters measuring different attributes or aspects of the same attribute.

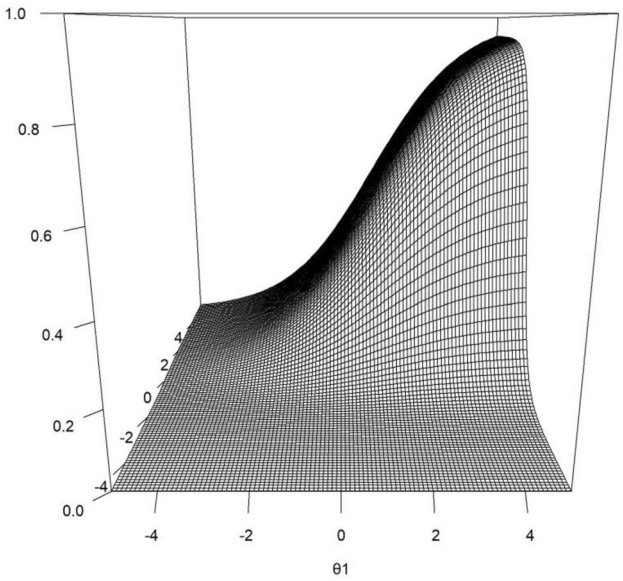


Fig. 2 The IRF of a dichotomous item in a two-dimensional latent-variable model. The first latent variable ($\theta 1$) is on the x-axis, the second latent variable ($\theta 2$; label not shown) is on the y-axis, and the probability of obtaining item score 1 given the values of $\theta 1$ and $\theta 2$ is on the z-axis

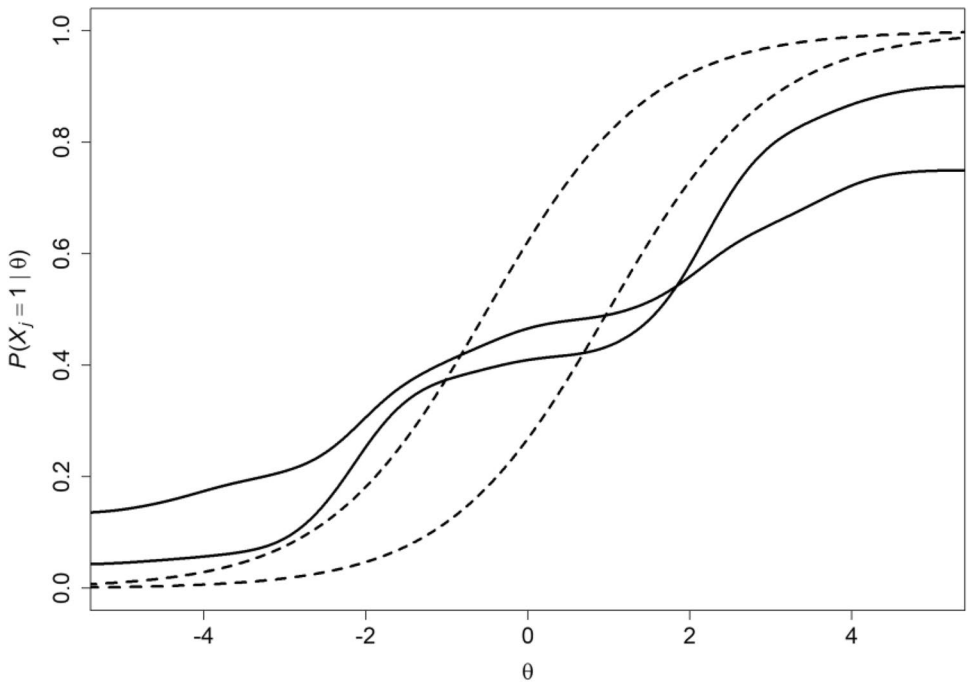
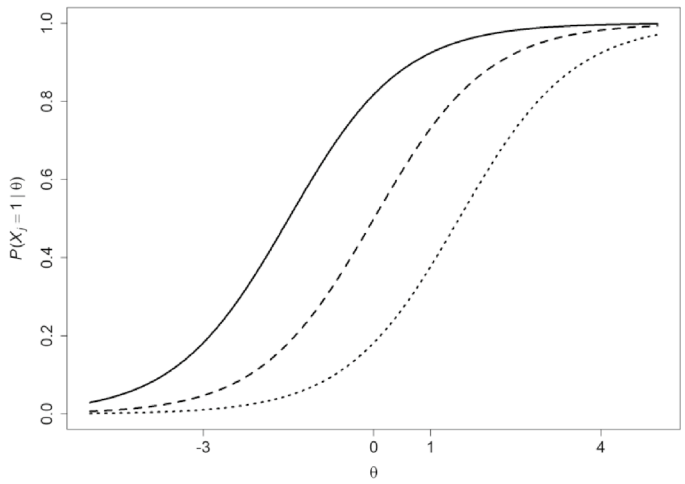


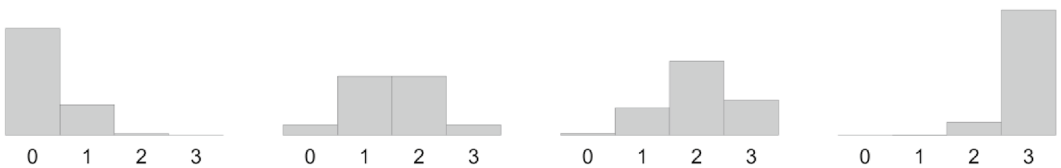
Fig. 1 Four items having monotone IRFs consistent with the monotone homogeneity model, of which two (dashed logistic curves) also follow the Rasch model

Other features render nonparametric IRT models interesting. First, for binary-item tests and questionnaires consistent with the model of monotone homogeneity, ordering persons by their number of positive scores (the total number of 1 scores or the sum score) is stochastically equal (i.e., with possible random violations) to ordering them by means of their latent-variable scores. This property says that one does not need the latent-variable scores for ordering persons and that simple sum scores will suffice albeit with random error (Fig. 3). This is a strong result, because it holds for scales when the model of monotone homogeneity fits the data well, and one need not estimate latent-variable scores at all. Because the Rasch model and the 2-parameter and 3-parameter logistic models are special cases of the model of monotone homogeneity, these parametric IRT models imply the ordering property using the sum score as well. Unlike nonparametric IRT models, parametric IRT models estimate the latent variable and assign θ estimates to persons. If the purpose is to order persons on a scale, it does not matter whether one uses the estimated θ or the sum score. Both scores are liable to unreliability due to random error, and may not perfectly reflect the error-free ordering of persons. Unlike nonparametric IRT models, parametric IRT models allow the assessment of scale-dependent measurement precision using the estimated θ. For polytomous-item tests and questionnaires, the ordering of persons by their sum scores approximates their ordering by latent-variable scores quite well, but may contain small but unimportant distortions.

Second, one might argue correctly that assuming logistic IRFs provides efficient item information by means of



$\theta = -3$ $\theta = 0$ $\theta = 1$ $\theta = 4$



$E(X_+|\theta = -3) = 0.24$ $E(X_+|\theta = 0) = 1.50$ $E(X_+|\theta = 1) = 2.03$ $E(X_+|\theta = 4) = 2.90$

$X_+ = 0$ $X_+ = 2$ $X_+ = 1$ $X_+ = 3$

Fig. 3 Upper panel: Three Rasch items (locations $\delta_1 = -1.5$, $\delta_2 = 0$, $\delta_3 = 1.5$) and four example θ-values ($\theta_1 = -3$; $\theta_2 = 0, \theta_3 = 1, \theta_4 = 4$) plotted on the horizontal axis. Three dichotomous items allow four sum scores: $X_+ = 0, 1, 2, 3$. Lower panel: Histograms showing the the sum-score distribution for each θ value, and the corresponding expected (i.e., mean) sum score, $E(X_+|\theta)$. Expected sum scores $E(X_+|\theta)$ have the same ordering as θ-values. Last line: Sum-score values X_+ obtained by randomly drawing from the histograms. Unreliability causes different orderings of X_+ and θ in this particular draw

estimated difficulty, discrimination and perhaps pseudo-guessing parameters, but the fact that nonparametric IRT models do not commit to specific parametric IRFs, instead estimating the whole function for each item from the data allows a complete picture of item response behavior for each item. This allows the researcher to see that the item only works well for people high on the scale of, for example, physical functioning, but not for the majority (Fig. 4, solid curve), or that the IRF only has a weak and irregular relation with physical functioning and is a candidate for replacement (Fig. 4, dashed curve). Researchers and scale developers want to know these things and make decisions about maintaining, deleting or replacing items from preliminary tests or questionnaires. Only having estimates of where an item is located or whether it distinguishes people well at a particular scale location is useful but knowing the complete picture has greater diagnostic value for item assessment. Estimating IRFs of course is liable to sampling error and involves several rather arbitrary decisions [11].

Third, the fact that nonparametric IRT refrains from assumptions about response behavior that may be mathematically convenient but unnecessarily restrictive from a psychological or health-related point of view—which content-driven theory predicts that IRFs are S-shaped or run parallel—renders it more laborious when assessing whether the model is consistent with the data. Various methods exist for assessing IRF monotonicity and local independence and exploring the dimensionality of an item set, and several run into problems related to combinatorial explosions and the curse of dimensionality. They often provide a wealth of detail about goodness of fit, but have trouble summarizing these details into discrete often binary conclusions such as keeping an item or deleting it from the test or questionnaire. Parametric IRT models have the advantage of concentrating on a few parameters per item rather than estimating whole curves but at the basis of several of their goodness-of-fit statistics lie the observable data features that nonparametric IRT models explore in detail but find hard to summarize. Whereas nonparametric IRT sometimes struggles with all the details it wishes to involve, the level of granularity of parametric IRT may be too low, occasionally missing interesting data features. Of course, these characterizations are somewhat exaggerated to make our points but help to understand the topic of this article.

Goodness of fit of models to data and robustness when models fail to fit

Sijtsma and Van der Ark [1] discussed a methodology in ten steps for investigating the goodness of fit of the nonparametric IRT models of monotone homogeneity and double monotonicity to the data collected with a (preliminary) test or questionnaire. The authors focus on the simple case of one test or questionnaire administered once to a sample from the population of interest. The first three steps concern aspects of data examination, in particular, recoding of item scores, handling of inadmissible and missing values and finally the identification and handling of outliers. These steps are useful in any scale analysis, not particular a nonparametric IRT analysis and we will skip them here. The next four steps numbered 4–7, concern scale identification and are highly

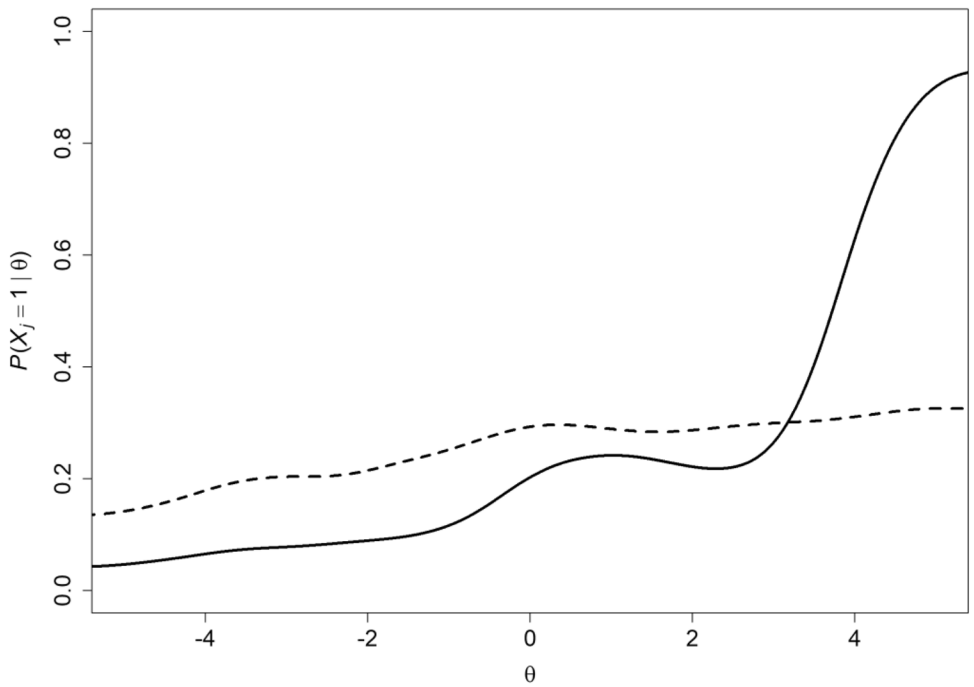


Fig. 4 Two IRFs showing one item that works well only for people with high latent-variable levels (solid curve), and one IRF that has a weak and irregular relation with the latent variable and is a candidate for replacement (dashed curve)

relevant in the nonparametric IRT context. The authors claim that these four steps together identify one or more scales that satisfy the model of monotone homogeneity and, if desirable, investigate whether the identified scales also are consistent with the more restrictive model of double monotonicity that requires the monotone IRFs not to intersect. Both models enable a person ordering using the simple sum score and the latter model also enables an item ordering by item means that is the same, with the exception of possible ties, across the scale of the latent variable. The last three steps again are relevant to any scale analysis model, not only nonparametric IRT, and are reliability estimation for the sum score, determining norms tables for the interpretation of sum scores of individuals, and the comparison of scaling results across meaningful subgroups of the population in interest. Due to their generality, we will also skip these last three steps.

We illustrate the steps briefly using data from the physical health questionnaire Climbing Stairs (PHQ-CS) [12]. The 15-item questionnaire was administered to 759 subjects with lower-extremity disorders living at home. Each item consists of a statement that describes a problem a patient might encounter when climbing stairs, and the respondent either endorsed the statement (score 1) or not (score 0). The first 12 items pertain to 6 aspects of stair climbing (takes longer, different way, with difficulty, hold on to banister, use walking aid, helped by someone) that are applied to going up and going down, respectively. The last 3 items pertain to the frequency of climbing stairs. The complete PHQ-CS is available from [12].

Step 4—scalability

A nonparametric IRT scale analysis usually starts with the determination of the dimensionality of the item set. That is, do we need one or more latent variables to explain the data structure and in case of multidimensionality, do the items divide neatly across two or more subclusters of items that are interpretable and form a basis for separate scales? Mokken [10] (also, see [13], Chap. 4) proposed an automated item selection algorithm based on scalability coefficient H producing one or more preliminary scales in which both the individual items and the total item (sub)set satisfy minimum requirements for H (called lower bounds). The requirements ascertain reliable person ordering on a scale spanned by the item (sub)set. Straat et al. [14] proposed an alternative item selection algorithm based on a genetic search. Brusco et al. [15], proposed an alternative clustering procedure. Zhang and Stout [16] discussed the DETECT procedure based on conditional covariances between item scores and Bolt [17] discussed related proposals. Van Abswoude et al. [18] used simulated data to compare various dimensionality assessment procedures. After determining a preliminary division of items in one or more item sets, for each set the assumptions of the nonparametric IRT models are investigated. In their contribution to the special section, Koopman et al. [19] discuss item selection based on scalability coefficients for clustered data common in much health research.

For the PHQ-CS, Mokken's [10] automated item selection procedure produced the same results for lower bounds in the range from 0.0 to 0.4 (Table 1, columns 3 and 4): Except item 4, all items formed a single scale. Item 4 was excluded due to a negative item-pair scalability coefficient with item 12 ($\widehat{H}_{4,12} = -.058$; SE = .417). Because deleting item 4 did not alter the scalability, and because of the small point estimate and the large standard error, we decided to maintain item 4 in the scale. The other estimated item scalability coefficients were all larger than the conventional lower bound 0.3. The estimated scalability of the entire scale was $\widehat{H} = .497$ (SE = .019). Following Mokken's [10] guidelines, $.4 < H \leq .5$ is a medium scale.

Step 5—local independence

To investigate the assumptions of nonparametric IRT models, one needs properties the models imply that do not contain the latent variable and can be computed directly from the data. An example is conditional association, and a special case of this property is the correlation between two item scores conditional on a function of the scores on one or more of the other items. Such a function is the sum score on these items, also called the rest score, which replaces the latent variable. These covariances must be nonnegative when the model of monotone homogeneity holds; negative values are inconsistent with the model. Straat et al. [20] used such conditional covariances to identify item pairs that are locally dependent, suggesting that their covariance does not only depend on the attribute one wishes to measure, but also on other, undesirable influences. Such items may be candidates for removal or replacement, or the researcher may maintain them when she assesses the inconsistency not serious enough. Also, see [21] for an approach based on nonparametric regression and the parametric bootstrap. For the PHQ-CS, using the method of conditional covariances [20] we detected several locally dependent item pairs (Table 1, last column). As this method was not investigated thoroughly, results should be interpreted with care [1].

Step 6—monotonicity

To investigate whether response probabilities of a positive answer or picking a particular response category or a higher one are monotone related to the latent variable, again we need to assess a function the model of monotone homogeneity implies that does not contain the latent variable and can be computed directly from the data. Such a

Table 1 Scaling results for PHQ-CS. Step 4 (Scalability) and Step 5 (Local Dependence): Automated item selection for lower bounds .0, .4, and .5; estimated item scalability coefficients ($\widehat{H}_j$) plus standard error (SE) for the scale consisting of all 15 items; overview of positive locally dependent (PLD) item pairs

Item	Statement	Lower bound[a]			$\widehat{H}_j$	SE	PLD item pairs[b]
		.0	.4	.5			
1	I go up the stairs but it takes longer	1	1	1	.618	(.024)	
2	I go up the stairs but in a different way	1	1	1	.447	(.025)	12
3	I go up the stairs but with (some) difficulty	1	1	2	.523	(.023)	
4	I go up the stairs and hold onto the banister			2	.576	(.039)	12
5	I go up the stairs and use a walking aid	1	1	1	.451	(.046)	11
6	I go up the stairs and am helped by someone	1	1	1	.418	(.097)	8, 10, 12
7	I go down the stairs but it takes longer	1	1	1	.578	(.023)	
8	I go down the stairs but in a different way	1	1	1	.437	(.026)	6, 12
9	I go down the stairs but with (some) difficulty	1	1	2	.527	(.023)	
10	I go down the stairs and hold onto the banister	1	1	1	.594	(.041)	6
11	I go down the stairs and use a walking aid	1	1	1	.506	(.044)	5
12	I go down the stairs and am helped by someone	1	1	1	.388	(.105)	2, 4, 6, 8
13	I do go up and down the stairs but less often	1	1	3	.419	(.026)	
14	I do go up and down the stairs but I avoid them	1	1	2	.447	(.027)	
15	I do go up and down the stairs but less stairs/ floors	1	1	3	.425	(.032)	

Columns 'Item' and 'Statement' adapted from "Measuring activity limitations in climbing stairs: development of a hierarchical scale for patients with lower-extremity disorders living at home", by Roorda et al. [12], Appendix. Copyright 2004 by Elsevier. Reprinted with permission

[a]1: item is selected into the first scale, 2: item is selected into the second scale; etc. Blank: item is unscalable

[b]If not a blank, the item may be in one or more positive locally dependent item pairs; the number indicates the other item in the positive locally dependent item pair(s)

function conditions the response probabilities for an item on the sum score based on the other items, which is the rest score we discussed previously and replaces the latent variable. This is the observable property called manifest monotonicity [22]. Various nonparametric regression methods were proposed for assessing the monotonicity assumption, based on binning [23], kernel smoothing [11, 24], and spline fitting [25, 26]. Local decreases in estimated IRFs suggest that the item is ineffective for measurement at those scale ranges, but model-consistent local increases that have an almost flat slope may also be informative.

Figure 5 shows the estimated IRFs of the PHQ-CS. The IRFs were estimated using kernel smoothing setting bandwidth $h = .08$ [11]. Using the banister is the most popular coping strategy, whereas using assistance is the least popular. The IRFs of items pertaining to the same aspect of ascending and descending are remarkably similar. Items 2, 5, 11, and 14 show minor violations of monotonicity.

In their contribution to the special section, Falk and Fischer [27] study a flexible approach based on monotonic polynomials that provides a compromise by modeling items with both complex and simpler response curves. Their study investigates the suitability of items with IRFs described by monotonic polynomials for inclusion in patient-reported outcomes item banks.

Step 7—invariant item ordering

The test is more informative when the ordering of the items by means of response probability or item mean score is the same, except for possible ties, for each measurement value. This means that if we know that for a particular scale value the probability of giving a positive response is greater for item j than item k, we know that this ordering is the same—but never opposite—for all other scale values. This is different in parametric IRT models, where one often takes the ordering of items' location parameters as their ordering according to difficulty or popularity, but when IRFs intersect, this is an incorrect conclusion (Fig. 6). Various methods exist for investigating whether sets of binary or polytomous items have an invariant ordering; see [28] for an automated methodology. Groundbreaking theoretical work was due to Rosenbaum [29], whereas Tijmstra et al. [30] provided new results.

Data analysis experience shows that an invariant item ordering is restrictive, rarely achieved for the whole set of items. For the PHQ-CS, an IIO is not achieved either,

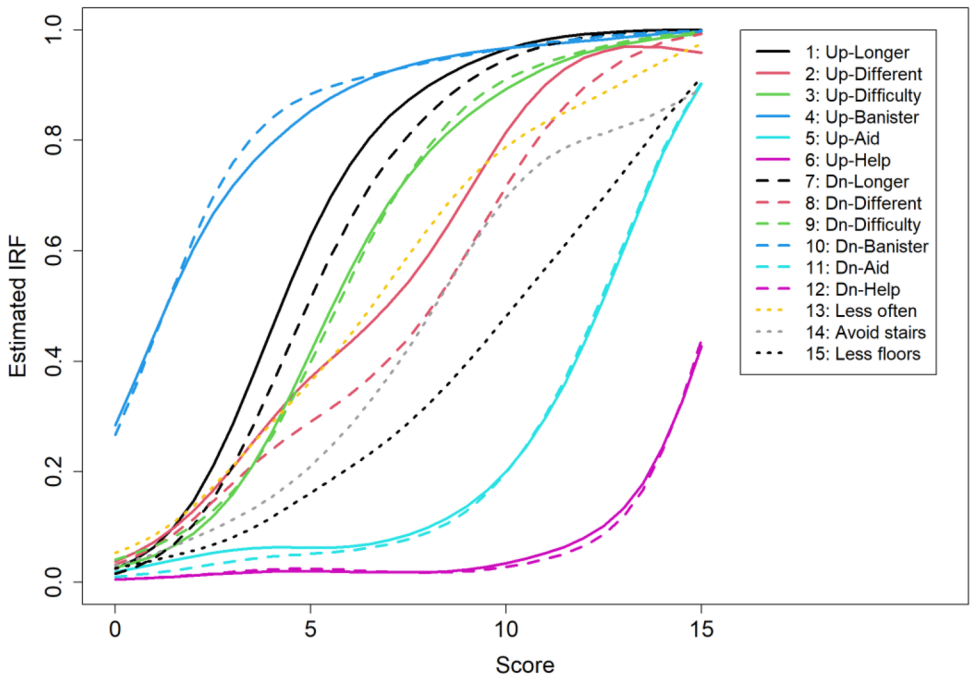


Fig. 5 Estimated IRFs of 15 items from the PHQ-CS [12]. See Table 1 for the full item content

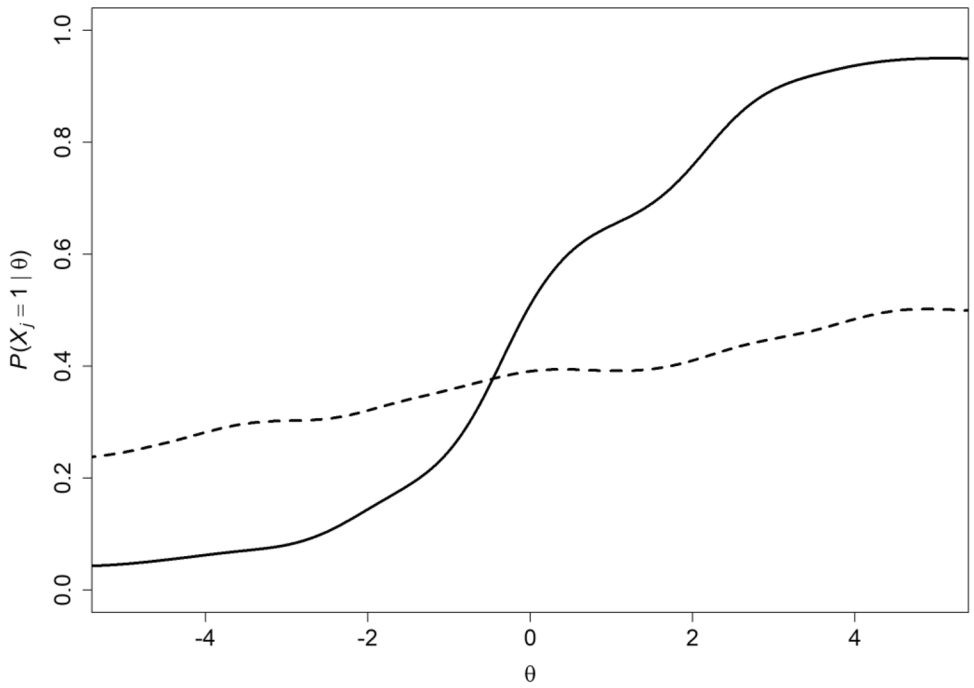


Fig. 6 Two intersecting IRFs

as the estimated IRFs in Fig. 5 cross. However, except for the item pair that relates to the aspect of 'climbing in a different way' (items 2 and 8), the pairs of items pertaining to the remaining five aspects seem to approximate an invariant ordering. 'Holding on to the banister' (items 4 and 10) is invariantly more popular than the remaining four aspects, 'taking longer' (items 1 and 7) is invariantly more popular than the remaining three aspects, 'climbing with difficulty' (items 3 and 9) is invariantly more popular than the remaining two aspects, and 'using a walking aid' (items 5 and 11) is invariantly more popular than 'being helped by someone' (items 6 and 12). The aspect 'climbing in a different way' is less popular than 'using walking aids' and more popular than 'climbing with difficulty' for the majority of patients, but not for the low-scoring patients.

When applied to real-data analysis, the previous four steps often provide detailed information on dimensionality, local

independence, monotonicity, and invariant item ordering, and the researcher finds herself confronted with the question how to weigh and combine the information to draw a conclusion about the scale(s). In this special section, Crisan et al. [31] critically discussed a summary measure called *crit* that was proposed as a heuristic tool to help researchers finding their way in the output of a Mokken Scale Analysis [23]. The four analysis steps pertain to the situation in which a test or questionnaire is administered once to a sample from the population of interest, but in repeated measurement, the issue of response shift—that is, a change in patients' perspective on the meaning of an item—may reveal itself through a change in item ordering at the individual level. In this special section, Dubuy et al. [32] mentioned chronic diseases where patients regularly adapt to their life circumstances, resulting in a different interpretation of items when tested repeatedly. In their contribution to the special section, these authors discuss a method to study this phenomenon of response shift for patient-reported outcomes.

Discussion

Nonparametric IRT scaling puts fewer constraints on the data than several parametric IRT models do. This way, nonparametric IRT retains a larger number of items from preliminary test and questionnaire versions, which not only is efficient but also provides a good fit to the state of theory development for attributes in social, psychological, and health sciences. That is, theories for attributes may predict an attribute as cumulative (e.g., intelligence) or categorical (e.g., typologies), but theories do not (yet) predict that response probabilities for different items run parallel or can be described sufficiently well with one, two, or three parameters. In all fairness, data analysis experience has taught us that the data structure at best only approximates unidimensionality, local independence, monotonicity, and invariant item ordering, and there always is at least some discrepancy between model predictions and data structure. The crucial issue with all modeling attempts is not whether the model fits the data, but whether the discrepancy of model and data is small enough for the model properties to hold for the application at hand. Much research addresses whether statistical tests provide a Type I error that is almost equal to the significance level the researcher choses, and how particular data features not anticipated by the model affect a statistical test's power. The question of the magnitude of discrepancy between IRT model and data is difficult to answer, because there are so many ways in which data can digress for the model's prediction. Moreover, the practical use of a test or a questionnaire determines the cost or utility of false positives and false negatives in relation to correct decisions based on the sum score. The five articles in this special section provide valuable psychometric contributions to the further development of measurement in the health sciences.

Funding The authors did not receive funding for this article.

Data availability The data are not publicly available.

Code availability Computer code is available from the R package mokken [33, 34].

Declarations

Conflict of interest The authors have no conflicts of interest to disclose.

Ethical approval This retrospective study involves the reanalysis of questionnaire data. The Institutional Review Board of Child Development and Education at the University of Amsterdam (number 2021-CDE-13947) granted ethical approval.

References

1. Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137–158. https://doi.org/10.1111/bmsp.12078
2. Van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory. Volume 1: Models*. Chapman & Hall/CRC.
3. Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes*. Chapman & Hall/CRC.
4. De Jong, A. E. E., Bremer, M., Schouten, M., Tuinebreijer, W. E., & Faber, A. W. (2005). Reliability and validity of the pain observation scale for young children and the visual analogue scale in children with burns. *Burns, 31*(2), 198–204. https://doi.org/10.1016/j.burns.2004.09.013
5. Gandek, B., Sinclair, S. J., Kosinski, M., & Ware, J. E., Jr. (2004). Psychometric evaluation of the SF-36® health survey in medicare managed care. *Health Care Financing Review, 25*(4), 5.
6. Ma, C., Chen, S., You, L., Luo, Z., & Xing, C. (2012). Development and psychometric evaluation of the Treatment Adherence Questionnaire for Patients with Hypertension. *Journal of Advanced Nursing, 68*(6), 1402–1413. https://doi.org/10.1111/j.1365-2648.2011.05835.x
7. Roorda, L. D., Molenaar, I. W., Lankhorst, G. J., Bouter, L. M. & Measuring Mobility Study Group. (2005). Improvement of a

questionnaire measuring activity limitations in rising and sitting down in patients with lower-extremity disorders living at home. *Archives of Physical Medicine and Rehabilitation, 86*(11), 2204–2210. https://doi.org/10.1016/j.apmr.2005.06.005
8. Sijtsma, K., Emons, W. H. M., Bouwmeester, S., Nykliček, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the World Health Organization Quality-of-Life scale (WHOQOL-Bref). *Quality of Life Research, 17*(2), 275–290. https://doi.org/10.1007/s11136-007-9281-6
9. Feng, Y.-S., Jiang, R., Pickard, A. S., & Kohlmann, T. (2021). Combining EQ-5D-5L items into a level summary score: Demonstrating feasibility using non-parametric item response theory using an international dataset. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-02922-1
10. Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton/De Gruyter.
11. Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*(3), 234–243. https://doi.org/10.1177/01466210122032046
12. Roorda, L. D., Roebroeck, M. E., Van Tilburg, T., Lankhorst, G. J., Bouter, L. M., & Measuring Mobility Study Group. (2004). Measuring activity limitations in climbing stairs: Development of a hierarchical scale for patients with lower-extremity disorders living at home. *Archives of Physiological and Medical Rehabilitation, 85*(6), 967–971. https://doi.org/10.1016/j.apmr.2003.11.018
13. Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.
14. Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*(1), 75–99. https://doi.org/10.1007/s00357-013-9122-y
15. Brusco, M. J., Köhn, H.-F., & Steinley, D. (2015). An exact method for partitioning dichotomous Items within the framework of the monotone homogeneity model. *Psychometrika, 80*(4), 949–967. https://doi.org/10.1007/s11336-015-9459-8
16. Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213–249. https://doi.org/10.1007/BF02294536
17. Bolt, D. M. (2001). Conditional covariance-based representation of multidimensional test structure. *Applied Psychological Measurement, 25*(3), 244–257. https://doi.org/10.1177/01466210122032055
18. Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*(1), 3–24. https://doi.org/10.1177/0146621603259277
19. Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2021). A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-02840-2
20. Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12*(4), 117–123. https://doi.org/10.1027/1614-2241/a000115
21. Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement, 25*(3), 221–233. https://doi.org/10.1177/01466210122032037
22. Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*(3), 1359–1378. https://doi.org/10.1214/aos/1176349262
23. Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for windows user's manual*. ProGAMMA.
24. Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611–630. https://doi.org/10.1007/BF02294494
25. Abrahamowicz, M., & Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika, 57*(1), 5–27. https://doi.org/10.1007/BF02294656
26. Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal and Educational and Behavioral Statistics, 27*(3), 291–317. https://doi.org/10.3102/10769986027003291
27. Falk, C. F., & Fischer, F. (2021). More flexible response functions for the PROMIS physical functioning item bank by application of a monotonic polynomial approach. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-02873-7
28. Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*(4), 578–595. https://doi.org/10.1177/0013164409355697
29. Rosenbaum, P. R. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40*(2), 157–168. https://doi.org/10.1111/j.2044-8317.1987.tb00875.x
30. Tijmstra, J., Hessen, D. J., Van der Heijden, P. G. M., & Sijtsma, K. (2011). Invariant ordering of item-total regressions. *Psychometrika, 76*(2), 217–227. https://doi.org/10.1007/s11336-011-9201-0
31. Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2021). The *Crit* coefficient in Mokken Scale Analysis: A simulation study and an application in quality-of-life research. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-02924-z
32. Dubuy, Y., Sébille, V., Grall-Bronnec, M., Challet-Bouju, G., Blanchin, M., & Hardouin, J.-B. (2021). Evaluation of the link between the Guttman errors and response shift at the individual level. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-03015-9
33. Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19. https://doi.org/10.18637/jss.v020.i11.
34. Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. https://doi.org/10.18637/jss.v048.i05.