Nonparametric item response theory for multilevel test data

Koopman, L.

Publication date
2022

Chapter 1

Introduction

1.1 Positioning of this Thesis

Psychological measurement aims at measuring an unobservable psychological construct (e.g., an attribute, skill, or ability) by means of observable variables such as items in a test or questionnaire (e.g., Sijtsma & Van der Ark, 2020, pp. 5–6). This type of measurement is omnipresent in many areas of the social and behavioral sciences, including education, clinical psychology, personality assessment, and (health-related) quality of life research. Common practice is to use the test score (e.g., the sum or mean score across the items in a test) to measure respondents' levels on the psychological construct. For example, one may claim that higher test scores on a questionnaire for extraversion represent higher levels of extraversion in respondents, or that students with a dyslexia test score exceeding a preset cut-off point have such severe reading difficulties they should receive learning support.

Item response theory (IRT) models are psychometric measurement models used to relate item scores to the psychological construct, in which the construct is represented by a (possibly multidimensional) latent variable (e.g., Bartholomew et al., 2011; Embretson & Reise, 2000; Sijtsma & Van der Ark, 2020; Van der Linden, 2016). Different IRT models imply different properties for the scale on which the construct is measured, such as whether the respondents can be ordered on the latent variable using the test score, or whether values on the latent variable can be estimated using the item scores. Hence, a fitting model informs psychometricians, substantive researchers, and other practitioners on the characteristics and properties of the test and its items. Because IRT models and its features are only valid if they fit test data well (Sijtsma & Van der Ark, 2020, p. 3), validating a test or questionnaire using psychometric measurement models is a vital step before using it for measurement (see, also, Borsboom et al., 2004).

IRT models vary in their degree of restrictiveness with regards to the structure of the latent variable and its relation to the items. *Nonparametric IRT* models put relatively few restrictions on the data, and therefore fit relatively well to test data, making them attractive measurement models. The scaling of items under nonparametric IRT models, and methods for investigating the fit, are collectively known as *Mokken scale analysis* (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017). The research of this thesis focuses on generalizing nonparametric IRT models and MSA to multilevel test data. In the remainder of this chapter, I first briefly discuss IRT and MSA. Next, I briefly discuss a nonparametric approach to multilevel IRT. Finally, I provide an outline of this thesis

1.2 Item Response Theory

Item response theory (IRT; Embretson & Reise, 2000; Sijtsma & Molenaar, 2002; Van der Linden, 2016) constitutes a family of psychometric measurement models that vary mainly with respect to how item scores are related to latent variables. This relation is formalized in an *item-response function*, which provides the expected item score given the value on latent variable(s). IRT models vary with respect to the restrictiveness in their model assumptions. However, most IRT models are defined by at least these three assumptions: Unidimensionality of the latent variable, local independence of the item scores given the value on latent variable, and monotonicity of the item-response function (i.e., the item-response function is nondecreasing across the latent variable).

A model consisting of only these three main IRT assumptions was introduced by Mokken (1971) as the *monotone homogeneity model* (MHM; Sijtsma & Molenaar, 2002; a.k.a. monotone unidimensional representation, Junker, 1993; Junker & Ellis, 1997; unidimensional monotone latent variable model, Holland & Rosenbaum, 1986; nonparametric graded response model, Hemker et al., 1996, 1997). The MHM is a nonparametric IRT model, because it makes no further distributional assumptions on the latent variable or the item-response function. Figure 1.1 shows three nonparametric item-response functions that comply with the MHM. A nonparametric special case of the MHM is the *double monotonicity model* (DMM; Mokken, 1971; Sijtsma & Molenaar, 2002), which additionally assumes an invariant item ordering (i.e., the item-response functions do not intersect; Ligtvoet et al., 2011; Sijtsma & Junker, 1996; Sijtsma & Hemker, 1998)[1]. Other nonparametric IRT models are the nonparametric partial credit model and the nonparametric sequential model (Hemker et al., 1997, 2001, respectively).

Nonparametric IRT (Mokken, 1971; Sijtsma & Molenaar, 2002) models strive to obtain useful measurement properties using as few restrictions as possible (Holland & Rosenbaum, 1986; Junker & Ellis, 1997; Rosenbaum, 1984). The MHM and the DMM imply ordering properties with respect to respondents and, for the DMM, items (e.g., Grayson, 1988; Hemker et al., 1997; Ligtvoet et al., 2011; Sijtsma & Junker, 1996; Sijtsma & Hemker, 1998; Van der Ark & Bergsma, 2010). These properties are useful for, for example, using

[1] For polytomous items, the DMM was originally defined using nonintersecting item-step response functions (i.e., all item-score categories are ordered), rather than nonintersecting item-response functions (Molenaar, 1997). As investigating properties of items can be considered more relevant than investigating properties of item-steps, the new definition of the DMM can be considered more useful (Sijtsma & Van der Ark, 2017, 2020, p. 158).

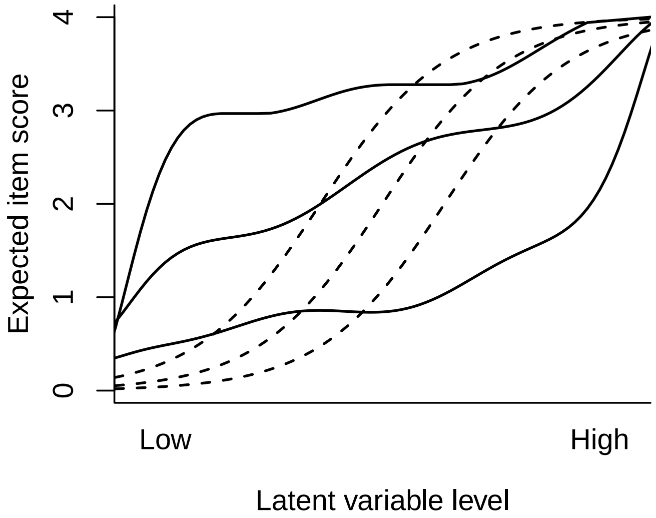

Figure 1.1: Three nonparametric item-response functions (solid lines) and three parametric item-response functions (dashed lines)

the test scores to order respondents from high to low on some ability, selecting the 30% most capable respondents for a special training program, constructing percentile scores, or presenting items on an intelligence test in ascending difficulty. In addition, the MHM and DMM pose various restrictions on the data that can be used to evaluate model fit, including manifest monotonicity (Junker, 1993; Sijtsma & Hemker, 2000), conditional association (Holland & Rosenbaum, 1986; Straat et al., 2016), modified scalability bounds (Ellis, 2014), and, for the DMM, manifest invariant item ordering (Ligtvoet et al., 2010).

Parametric IRT models place distributional assumptions on the latent variable and/or the item-response function. All popular unidimensional parametric IRT models, such as the Rasch Model (Rasch, 1960), the two- and three-parameter logistic models (Birnbaum, 1968), the graded response model (Samejima, 1969), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), and the sequential model (Tutz, 1990) are special cases of the MHM (Van der Ark, 2001). Because parametric IRT models are special cases of the MHM, they imply the same properties, and in addition enable other properties. For example, the Rasch model implies that respondent and item parameters can be estimated on the latent variable, independent of the items or respondents involved, respectively (Van der Linden, 2016, p. 33). This is useful for, for example, computerized adaptive testing (Luijten et al., 2021; N. Smits et al., 2018; Van der Linden & Glas, 2010). However, parametric model assumptions are more restrictive compared to nonparametric IRT models, and as a result, items are less likely to satisfy parametric model assumptions. Figure 1.1 (dashed lines) shows three item-response functions that comply with the Rasch model. Because the Rasch model is a special case of the MHM, the item-response functions of the Rasch model also satisfy the assumptions of the item-response-functions of the MHM.

Nonparametric IRT is favored over parametric IRT in various cases (see, e.g., Mokken, 1971, pp. 115–116; Sijtsma & Van der Ark, 2020, pp. 107–109). First, for measurement applications in which the test score is used to order respondents, a fitting nonparametric IRT model is sufficient whereas the more restrictive parametric IRT models need not fit the data. Second, there may be limited information about newly constructed items and their relation to the latent variable. Nonparametric IRT models pose no restrictions on this relation beyond monotonicity, providing an opportunity to investigate item characteristics using methods in MSA. If desired, this may serve as a preliminary analysis to parametric IRT modeling, because if a nonparametric IRT model does not fit, neither does a parametric special case. Third, for psychological constructs for which creating a large number of items is difficult, retaining as many items as possible is desired. Using a nonparametric IRT modelling approach maintains items in a test that contribute to accurate measurement, even if they do not comply with the mathematical restrictions required by parametric IRT models, resulting in larger scales that have the stochastic ordering property.

1.3 Mokken Scale Analysis

Mokken scale analysis (MSA) consists of several procedures for constructing and evaluating tests and questionnaires (or *scales*) based on nonparametric IRT models (see, e.g., Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017). Generally, a scale is evaluated or constructed based on scalability coefficients, which are instrumental for determining the degree to which a set of items form a single scale (Mokken, 1971, p. 174). Scalability coefficients exist for item pairs (H_{ij}), for items (H_i), and for the entire test (H). A Mokken scale is defined as a scale for which the following two criteria hold:

- $H_{ij} > 0$ for all item pairs
- $H_i \geq c$ for all items, with c some positive constant

Using these two criteria, a fixed item-set may be evaluated, or scales may be constructed via an *automated item selection procedure* (AISP) by subsequently adding items to a scale as long as the criteria of a Mokken scale are satisfied.

Scalability coefficients are based on the number of Guttman errors in the data. A Guttman error is defined as passing a more difficult (or less popular) item after failing an easier (or more popular) item (Guttman, 1950). For example, in general we may expect that the item *"I don't mind working on mathematical problems"* is more popular than the item *"I enjoy working on mathematical problems"*. In this case, not endorsing the first item but endorsing the second item would constitute a Guttman error within respondents. If there are zero Guttman errors in the data, the scalability coefficient of this item pair has a value of 1. If there are as many Guttman errors as would be expected under independence of the items, this scalability coefficient has a value of 0. Guttman errors may also be counted within a particular response pattern to detect aberrant response patterns or outliers (Conijn et al., 2020; Meijer, 1994).

After establishing the scalability of the items, model fit methods may be applied for further inspection of the items. For example, conditional association may provide information on violations of the local independence assumption (Holland & Rosenbaum, 1986; Straat et al., 2016), manifest monotonicity on violations on the monotonicity assumption (Junker, 1993), and manifest invariant item ordering on violations of the invariant item ordering assumption (Ligtvoet et al., 2010). Results from these analyses, along with theoretical and content considerations, guide conclusions pertaining which items to include in the final scale (Crişan et al., 2020).

1.4 Multilevel Item Response Theory

Multilevel test data, consisting of item scores of respondents nested in clusters, are common in the social and behavioural sciences; for example item scores of students nested in classes, inhabitants nested in neighborhoods, employees nested in companies, or patients nested in hospitals. Item scores may pertain either to level 1 (respondent level) or level 2 (group level). When students who are nested in classes respond to the item "I enjoy explaining mathematics to others", the resulting item scores pertain to the students at respondent level. When students who are nested in classes respond to the item "The teacher seems to enjoy explaining mathematics to the class", the resulting item scores pertain to the teachers at the group level. If the respondent is of primary interest, the type of data may also be referred to as clustered data. If the grouping variable is of primary interest, the respondents may be considered the raters of the group, and this type of data may also be also referred to as multi-rater data. Statistically, there is no difference between the two types of test data: Item scores are dependent on the respondent level and the group level. Hence, we refer to this type of data in general as multilevel test data. The level of interest is relevant, however, to decide which assumptions are required to make justifiable conclusions on the test scores and, consequently, which analyses are informative. Therefore, we explicitly make the distinction between the data types.

There are two main reasons why currently MSA is not suitable for multilevel test data. First, methods in MSA assume that item scores are obtained using simple random sampling; that is, the respondents in the sample must not be clustered. In multilevel data this assumption is violated. Ignoring a nested structure is a well-known cause of underestimated standard errors (e.g., Maas & Hox, 2005; Hox, 2010). As a result, confidence intervals will be too narrow, making the estimates appear more precise than they actually are. In addition, the type I error rates of significance tests will be inflated, meaning that hypotheses are too often rejected. This may lead to the inclusion of items that do not contribute to (or possibly negatively affect) accurate measurement to the scale, or the strength of the scale may be overestimated. Second, most methods in MSA provide results on the respondent level only, not on the group level. Interpreting respondent-level results as if they are group-level results may falsely suggest satisfactory reliability and scalability coefficients for a set of items (Crişan et al., 2016). Hence, for scaling groups, the current available MSA methods are of limited value.

The research in thesis aims at developing nonparametric IRT and MSA for multilevel test data. In particular, this thesis builds on the work of Snijders (2001a), who proposed a two-level nonparametric IRT model for scaling groups scored by multiple respondents (see also Crişan et al., 2016; Reise et al., 2006). The model assumes unidimensionality, local independence, and monotonicity at the level of the respondent, and furthermore that the respondents are independent given the group. In addition, Snijders defined *two-level scalability coefficients* that provide information on scalability of items on the respondent level (within-rater scalability coefficients) and the group level (between-rater scalability coefficients). Within-rater scalability coefficients are similar to Mokken's scalability coefficients, based on Guttman errors within respondents. Between-rater scalability coefficients are based on Guttman errors between respondents within the same group. For example, if item 1 is more popular than item 2, a Guttman error between respondents happens if one respondent in the group does not endorse item 1, whereas another respondent endorses item 2. We use Snijders' model as a starting point for generalizing nonparametric IRT models and MSA methods to the more general framework of multilevel test data; that is, we develop models and methods for situations where the main focus is on the respondent level and for situations where the main focus is on the group level. Note that several parametric IRT models for multilevel test data have been proposed, such as the hierarchical rater model (Patz et al., 2002), the multiple raters model (Verhelst & Verstralen, 2001), and the rater bundle model (Wilson & Hoskens, 2001). In addition, Bayesian item response modeling can estimate a wide range of parametric IRT models in multilevel test data (Fox, 2010, Chapter 6).

1.5 Thesis Outline

The research presented in this thesis step-by-step develops MSA for multilevel test data, and in the process solves some computational and methodological issues that exist in traditional MSA. Chapter 2 introduces Guttman errors and solutions to two computational issues pertaining estimating Guttman errors. In Chapter 3, standard errors were derived using a marginal modeling framework and the delta method for all two-level scalability coefficients. As a result, the precision of estimated scalability coefficients can be determined, leading to more information with respect to the scalability of the items and the total test. Chapters 4 to 6 focus on evaluating and optimizing point and interval estimates of two-level scalability coefficients. Chapter 4 presents a Monte Carlo simulation study in which bias of estimated two-level scalability coefficients, the bias of their estimated standard errors (with two estimation methods), and the coverage of the confidence intervals were investigated, under various testing conditions. Chapter 5 provides a transformation for the scalability coefficients and their standard errors, particularly useful for very strong scales. Scalability coefficients take values on the interval $(-\infty, 1]$. The sampling distribution of scalability coefficients is skewed near the boundary, so Wald-based confidence intervals and significance tests may be biased. The transformation can be used to construct range-preserving confidence intervals and significance tests. In Chapter 6, I derived point estimates, standard errors, and test statistics for scalability coefficients in clustered data. In addition, I conducted a Monte Carlo simulation study that compared traditional estimation methods to these derived estimation methods, and compared Wald-based methods to range-preserving methods from Chapter 5 for scalability coefficients in clustered data. In Chapter 7, I used the results from Chapter 5 and 6 to incorporate significance tests for both criteria of a Mokken scale in the AISP algorithm, to take sample fluctuations into account. The result was a test-guided AISP, which was integrated into a two-step, test-guided MSA for scale construction, to guide the analysis for nonclustered and clustered data. In Chapter 8, I introduced four two-level nonparametric IRT models, their assumptions, the implied stochastic ordering properties, and the implied observable data properties that are useful for model fit investigation. In addition, I also derived the relations between models and properties. Finally, Chapter 9 is a general discussion of the results from the previous chapters, with guidelines for practical use, and suggestions for future research.