This is a “preproof” accepted article for *Psychometrika*.
This version may be subject to change during the production process.
DOI: 10.1017/psy.2024.9

Assumptions and Properties of Two-Level Nonparametric Item Response Theory Models

- Letty Koopman, Groningen Institute for Education and Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands.
  E-mail: L.Koopman@RUG.nl.
- Bonne Zijlstra, Research Institute of Child Development and Education, University of Amsterdam, P. O. Box 15776, 1001 NG Amsterdam, The Netherlands.
  E-mail: B.J.H.Zijlstra@UvA.nl.
- Andries van der Ark, Research Institute of Child Development and Education, University of Amsterdam, P. O. Box 15776, 1001 NG Amsterdam, The Netherlands.
  E-mail: L.A.vanderArk@UvA.nl.

Corresponding author:

Letty Koopman, Groningen Institute for Education and Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands. E-mail: L.Koopman@RUG.nl.

Declarations / Compliance with Ethical Standards:

Funding: This study was funded by the Netherlands Organisation for Scientic Research (NWO) (grant number 406.16.554).
Conflict of Interest: We have no conflicts of interest to disclose.
Ethical approval: This article does not contain any studies with human participants performed by any of the authors.
Availability of data, material, and code: No data, material, and code were used to prepare this manuscript.
Competing Interests: The authors declare none.

Abstract

Nonparametric item response theory (IRT) models consist of assumptions that restrict the joint item-score distribution. These assumptions imply stochastic ordering properties that allow ordering of respondents and items using the simple sum score and item mean score, respectively, and imply observable data properties that are useful for investigating model fit. In this paper, we investigate these properties for two-level nonparametric IRT. We introduce four two-level nonparametric IRT models. Two models pertain to respondents nested in groups: The MHM-1, useful for ordering respondents and groups, and the DMM-1, useful for ordering respondents, groups, and items. Two models pertain to groups rated by multiple respondents: The MHM-2, useful for ordering groups, and the DMM-2, useful for ordering groups and items. We define the model assumptions, derive implied stochastic ordering properties, and derive observable data properties that are useful for model fit investigation. Relations between models and properties are also presented.

Key words: conditional association, latent variable models, manifest invariant item ordering, manifest monotonicity, nonparametric item response theory, stochastic ordering.

1 Introduction

Most item response theory (IRT) models implicitly assume that the respondents are a random sample from the population envisaged. These IRT models assume one or possibly more latent variables only at the level of the respondent, and we refer to these IRT models as single-level IRT models. However, in many practical situations the respondents are nested in groups. For example, students nested in school classes

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

rating their teacher's instructional quality (Scherer et al., 2016), employees of the same department assessing humor in the workplace climate (Cann et al., 2014), or nurses within the same intensive care unit evaluating collaboration (Dougherty & Larson, 2010). In such situations, it is inappropriate to assume that the respondents are a random sample due to the group effect. It is therefore reasonable to use IRT models with a latent variable both on the respondent level and the group level (e.g., De Jong & Steenkamp, 2010; Fox, 2007; Fox & Glas, 2001). We refer to these IRT models as two-level IRT models. This paper investigates the measurement properties of a general nonparametric two-level IRT model, which was proposed by Snijders and Bosker (2012), and which can be considered a two-level generalization of the single-level non-parametric IRT models proposed by Mokken (1969) and Holland and Rosenbaum (1986).

Assume that a test consists of I items, indexed by i ($i = 1, 2, \ldots, I$), and each item has $m+1$ ordered item scores $0, 1, \ldots, m$. Assume that this test is administered to R randomly selected non-nested respondents, indexed by r ($r = 1, \ldots, R$). Note that index r refers to the rth respondent in the sample. Before sampling, it is not known which respondent from the population will be the rth respondent in the sample. Therefore, X_{ri} — defined as the score of the randomly selected rth respondent in the sample on item i — is a random variable. In this paper, variables will be denoted by uppercase letters, and their realizations by lower case letters. Hence, the realization of X_{ri} is denoted by x_{ri}. For each respondent, the I item scores can be collected in a vector $\mathbf{X}_r = (X_{r1}, X_{r2}, \ldots, X_{rI})$. Because the respondents are randomly and independently sampled, we consider the R vectors $\mathbf{X}_r$ independent and identically distributed (i.i.d.) for all r. As the respondents are non-nested, a single-level IRT model may be appropriate as a measurement model. Let Θ_r be a random latent variable of the rth randomly sampled respondent. Analogous to X_{ri}, Θ_r is a random variable, because before sampling it is not known which respondent from the population will be the rth respondent in the sample. Because the respondents are randomly and independently sampled, the R variables Θ_r are i.i.d. for all r. Let θ_r be a value of respondent r on the random latent variable Θ_r. For respondent r, the expected value on item i is $E(X_{ri}|\Theta_r = \theta_r) = \sum_{x=1}^{m} P(X_{ri} \geq x|\Theta_r = \theta_r)$. The expectation of X_{ri} as a function of Θ_r, $E(X_{ri}|\Theta_r)$, is referred to as the item response function (IRF; Chang & Mazzeo, 1994). Most single-level IRT models are defined by at least these three assumptions:

1. Unidimensionality (UN): Latent variable Θ_r is unidimensional

2. Local independence (LI): Item scores X_{ri} are independent given θ_r

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

3. Monotonicity (MO): $P(X_{ri} \geq x|\Theta_r = \theta_r)$ is nondecreasing in θ_r, for all i and for $x = 1, \ldots, m$

These assumptions are necessary to restrict the distribution of $\mathbf{X}_r$ (Junker & Ellis, 1997). The combination of UN, LI, and MO is also referred to as the *monotone homogeneity model* (MHM, Mokken, 1971; Sijtsma & Molenaar, 2002; a.k.a. monotone unidimensional representation, Junker, 1993; Junker & Ellis, 1997; unidimensional monotone latent variable model, Holland & Rosenbaum, 1986; and nonparametric graded response model, Hemker et al., 1996, 1997). The MHM does not use parameters to model the distribution of Θ and the relation between the item scores and Θ_r. The MHM is therefore called a nonparametric IRT model.

A fourth assumption in nonparametric IRT is invariant item ordering. Suppose that the I items are ordered by mean item score and numbered accordingly; that is, if $i < j$, then $E(X_{ri}) \leq E(X_{rj})$ for all $i \neq j$. Then,

4. Invariant item ordering (IIO): $E(X_{ri}|\Theta_r = \theta_r) \leq E(X_{rj}|\Theta_r = \theta_r)$ for all θ_r

(Ligtvoet et al., 2011; Sijtsma & Hemker, 1998; Sijtsma & Junker, 1996). IIO means that the order in difficulty is identical across all values of the latent variable. IIO allows the stochastic ordering of the items using the mean item scores. For applications of IIO we refer to Sijtsma et al. (2011). Following Sijtsma and Van der Ark (2017, 2020, pp. 156–158; also see the Discussion), we call the model that assumes UN, LI, MO, and IIO the double monotonicity model (DMM).

The MHM has several ordering properties. The MHM implies stochastic ordering of the manifest variable by the latent variable (Hemker et al., 1996, 1997), which implies that latent variable can be used stochastically to order the respondents on the unweighted sum score. More importantly, for dichotomous items, the MHM implies monotone likelihood ratio (MLR; Grayson, 1988; Huynh, 1994; Ünlü, 2008), which implies the property of stochastic ordering of the latent variable by the sum score across the items (SOL; Hemker et al., 1997). Measurement properties MLR and SOL imply that the sum score can be used to (stochastically) order respondents on the latent variable. For polytomous items, the MHM does not imply MLR and SOL (Hemker et al., 1996, 1997); however, the MHM implies the measurement property of weak SOL (Van der Ark & Bergsma, 2010), which can be used for pairwise ordering of respondents or groups on the latent variable.

These theoretical results justify ordinal person measurement by means of sum score X_{r+} if the MHM holds. Suppose that two respondents have sum scores a and b, respectively ($a < b$), then for dichotomous items, due to the SOL property, the MHM implies $E(\Theta_r|X_{r+} = a) \leq E(\Theta_r|X_{r+} = b)$; for polytomous items, due to the weak

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

SOL property, the MHM implies $E(\Theta_r|X_{r+} < a) \leq E(\Theta_r|X_{r+} \geq a)$. Hence, the sum score stochastically orders the respondents on Θ_r. Alternatively, suppose that two respondents have latent variable values t and u, respectively ($t < u$), then due to the monotonicity assumption the MHM implies $E(X_{r+}|\Theta_r = t) \leq (X_{r+}|\Theta_r = u)$. Hence the latent variable values stochastically orders the respondents on the sum score, a property sometimes referred to a stochastic ordering of the manifest variable by Θ_r (SOM; Hemker et al., 1997). These mutual ordering properties of X_{r+} and Θ_r, make X_{r+} an attractive estimator of Θ_r. Under the MHM, X_{r+} is a consistent asymptotic normal estimator of Θ_r (Junker, 1991; Stout, 1990).

The simple sum score is more intuitive for non-psychometricians than, for example, an estimated latent variable, because the sum score is defined on the scale of the test. Therefore, a higher sum score has a fairly straightforward interpretation, such as responded to more items correctly or responded more extreme to the items (Sijtsma & Hemker, 2000). In addition, using the sum score in scientific research avoids sample-specific transformations, which benefits comparability across studies and contributes to the replicability of results across studies (Edelsbrunner, 2022; Widaman & Revelle, 2022). Hence, providing justification for using the sum score is relevant for psychometric research and testing practice, even when the estimated latent variable is used for test construction and measurement evaluation (Hemker et al., 2001).

The DMM implies an ordinal scale for both person and item measurement. Hence, besides using the respondent sum score to order respondents on a latent variable, the mean item score can be used to order the items on a latent difficulty scale. Using the mean item score has similar advantages as the sum score for psychometric and testing practice: They have an intuitive interpretation, such as the proportion correct or average extremeness in the sample. In addition, estimating a latent difficulty is not straightforward and can have various interpretations that do not necessarily relate to the difficulty in practice (Sijtsma & Hemker, 2000; Sijtsma & Meijer, 2001).

All popular unidimensional 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). Hence, if the goal of the test is to order respondents, the MHM is preferred over popular parametric IRT models because, by definition, the MHM fits better to the data than these parametric IRT models. If the goal of the test is estimating the respondents' scores on Θ_r, alternative methods are required, such

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

as a smoothing procedure or estimating a parametric IRT model (e.g., Ramsay, 1991; Sijtsma & Van der Ark, 2020, Chapter 4, respectively). However, as these parametric IRT models are a special case of the MHM, investigating the fit of the MHM is still useful because if the MHM does not fit, neither do the parametric IRT models.

The MHM poses testable restrictions on the data, referred to as *observable properties*. For example, the MHM implies non-negative inter-item covariances (e.g., Sijtsma & Molenaar, 2002, pp. 155–156). Observable properties can be investigated in data to find evidence against the MHM assumptions. Holland and Rosenbaum (1986) showed that the MHM implies *conditional association* (CA). Let $\mathbf{Y}_r$ and $\mathbf{Z}_r$ be two mutually exclusive and exhaustive subsets of $\mathbf{X}_r$. CA holds if for every partitioning $\mathbf{X}_r = (\mathbf{Y}_r, \mathbf{Z}_r)$ and for all functions h, and for all non-decreasing functions g_1 and g_2

$$Cov[g_1(\mathbf{Y}_r), g_2(\mathbf{Y}_r)|h(\mathbf{Z}_r)] \geq 0. \tag{1}$$

The observable property CA is too comprohensive for a single testing procedure (see Ellis & Sijtsma, 2023) but special cases of CA, including testing for non-negative covariances, have been proposed to test the MHM. We focus on *manifest monotonicity* (MM Sijtsma & Hemker, 2000) and a testing procedure to identify locally dependent item sets using three cases of CA (Straat et al., 2016). For dichotomous items, CA implies MM (Ligtvoet, 2022). Let $X_{r(i)} = \sum_{j \neq i}^{I} X_{ri}$ be the rest score of item i, then MM means that

$$E(X_{ri}|X_{r(i)}) \text{ is nondecreasing in } X_{r(i)}. \tag{2}$$

Hence, MM is the MO assumption with latent variable Θ_r replaced by an observable proxy $X_{r(i)}$. Note that for polytomous items, the MHM does not imply MM. Straat et al. proposed testing $Cov(X_{ri}, X_{rj}) \geq 0$, $Cov(X_{ri}, X_{rj}|X_{rk}) \geq 0$, and $Cov(X_{ri}, X_{rj}|X_{r(ij)}) \geq 0$, where $X_{r(ij)} = \sum_{k \neq i,j} X_{rk}$. We refer to these three inequalities as *non-negative inter-item covariances* (NNIIC). As these three inequalities of NNIIC are special cases of CA with $g_1(\mathbf{Y}_r) = X_{ri}$, with $g_2(\mathbf{Y}_r) = X_{rj}$, and with $h(\mathbf{Z}_r) = \emptyset$, $h(\mathbf{Z}_r) = X_{rk}$, and $h(\mathbf{Z}_r) = X_{r(ij)}$, respectively, the MHM implies the three inequalities. Other CA-based observable properties have been proposed by, for example, Ellis (2014) and Ligtvoet (2022). Ellis and Sijtsma (2023) noted that these CA-based observable properties cannot distinguish between unidimensional and multidimensional models, and these authors suggested using (also CA-based) conditioning on added regression predictions (CARP) inequalities to investigate UN.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

The DMM poses additional observable properties (see Ligtvoet et al., 2011). We focus on *manifest invariant item ordering* (MIIO), which holds if for $E(X_{ri}) < E(X_{rj})$,

$$E(X_{ri}|X_{r(ij)} = y) \leq E(X_{rj}|X_{r(ij)} = y) \text{ for all } y \text{ and all } i < j. \tag{3}$$

Note that MIIO is the IIO assumption with latent variable Θ_r replaced by $X_{r(ij)} = \sum_{k \neq i,j} X_{rk}$. Other observable properties of IIO have been proposed; for example, by Tijmstra et al. (2011).

The assumptions (UD, LI, MO, IIO) discussed in this paragraph have not been formally defined for two-level IRT models, and as a result it is also unknown how these assumptions should be investigated in test data. Also, the measurement properties MLR, SOL, and SOM nor the observable properties MM, CA and MIIO have been defined for two-level IRT models, and as a result it is unknown whether two-level IRT models imply these measurement properties in the same way as single-level IRT models do. In the remainder of this paper, we generalize the MHM and DMM to two-level data on both the respondent level and the group level. We build on the work of Snijders (2001), who proposed a two-level nonparametric IRT model for scaling subjects (e.g., persons or groups) scored by multiple respondents (i.e., multi-rater measurement) using dichotomous items. For the proposed models, we establish which stochastic ordering properties and observable data properties are implied, and how they are related. Note that the proofs have been diverted to the Appendix. Implications and recommendations for practice and further research are discussed.

2 Two-Level Nonparametric IRT

Suppose a measurement instrument consists of I items, indexed by i or j ($i, j = 1, 2, \ldots, I; j \neq i$). Suppose there are S groups, indexed by s ($s = 1, 2, \ldots, S$), each consisting of R_s respondents, indexed by r ($r = 1, 2, \ldots, R_s$). Note that index s refers to the sth group and index r refers to the rth respondent in group s. Before sampling, it is not known which group from the population of groups will be the sth group in the sample, nor which respondent from the population of respondents will be the rth respondent in group s. The groups are assumed to be a random sample from a population of groups, and the respondents within a group are assumed to be a random sample from a population of respondents. Without loss of generality, we assume the number of respondents per group is the same; that is, $R_1 = R_2 = \cdots = R_S = R$. Let X_{sri} denote the score on item i of respondent r in group s,

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

with realization x_{sri} ($x_{sri} \in 0, \ldots, m$). For dichotomous items, $m = 1$ and x_{sri} takes on value 1 if item i is endorsed or answered correctly by respondent r in group s, and 0 otherwise. Let $X_{sr+} = \sum_{i=1}^{I} X_{sri}$ denote the respondent-level sum score. Let $X_{si} = R^{-1} \sum_{r=1}^{R} X_{sri}$ denote the group-level score on item i (i.e., the mean score over respondents' scores on item i within group s), with realization x_{si}. X_{si} can take on $Rm + 1$ values with a minimum of 0 and a maximum of m. Let $X_{s+} = \sum_{i=1}^{I} X_{si}$ denote the group-level sum score. The vector of item scores for respondent r in group s is denoted $\mathbf{X}_{sr} = (X_{sr1}, \ldots, X_{srI})$, with realization $\mathbf{x}_{sr} = (x_{sr1}, \ldots, x_{srI})$. Because the respondents within a group are randomly and independently sampled, the R vectors $\mathbf{X}_{sr}$ are considered i.i.d. within each s for all r. The vector of all item scores for group s is denoted $\mathbf{X}_s = (\mathbf{X}_{s1}, \ldots, \mathbf{X}_{sR}) = (X_{sr1}, \ldots, X_{sRI})$, with realization $\mathbf{x}_s = (\mathbf{x}_{s1}, \ldots, \mathbf{x}_{sR}) = (x_{sr1}, \ldots, x_{sRI})$. Because the groups are randomly and independently sampled, the S vectors $\mathbf{X}_s$ are considered i.i.d. for all s.

Let Θ_{sr}, Γ_s Δ_{sr} be random latent variables of the rth randomly sampled respondent in the sth randomly sampled group. Analogous to Θ_r in the single level situation, these are random variables because before the groups and respondents have been sampled, it is unknown which group from the population groups will be the sth group, and which respondent from the population of respondents belonging to the sth group will be the rth respondent. Variable Γ_s is considered a common group component, Δ_{sr} is a combination of an individual (random) respondent effect and a group by respondent interaction effect, and Θ_{sr} is the sum of these effects; that is,

$$\Theta_{sr} = \Gamma_s + \Delta_{sr}, \tag{4}$$

(Snijders, 2001). Let ε_{sri} be a random latent variable that may be interpreted as an error term. Assumption B is a basic assumption about the relation between the latent variables and the observed score X_{sri} using function f_i.

Assumption 1. *Basic assumption of item scores and latent variables.*

(B) $X_{sri} = f_i(\Gamma_s + \Delta_{sr}, \varepsilon_{sri})$. *For all s, r, and i, Γ_s, Δ_{sr}, and ε_{sri} are independent. Furthermore, all Γ_s ($s = 1, \ldots, S$) are identically distributed, and all Δ_{sr} ($s = 1, \ldots, S; r = 1, \ldots, R$) are identically distributed, with $E(\Delta_{sr}) = 0$.*

It follows from B that Θ_{sr} are identically distributed for all s, r, and that for a fixed item i, all ε_{sri} are identically distributed for all s, r. Assumption B is assumed throughout this paper. The variances of Θ_{sr}, Γ_s, and Δ_{sr} are denoted $\text{var}(\Theta_{sr})$, $\text{var}(\Gamma_s)$, and $\text{var}(\Delta_{sr})$, respectively. Because Γ_s and Δ_{sr} are assumed independent,

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

$\text{var}(\Theta_{sr}) = \text{var}(\Gamma_s) + \text{var}(\Delta_{sr})$, for all s and all r. Let θ_{sr} be a group-respondent combination value on Θ_{sr} of respondent r in group s, γ_s a value on Γ_s for group s, and δ_{sr} a value on Δ_{sr} for respondent r in group s. Hence, for respondent r in group s, we assume there exist value $\theta_{sr} = \gamma_s + \delta_{sr}$.

Let $P(\mathbf{X}_{sr} = \mathbf{x}|\Gamma_s, \Delta_{sr})$ denote the probability of obtaining item-score pattern $\mathbf{x}$ given Γ_s and Δ_{sr}. Throughout the rest of the paper we assume homogeneity of Γ_s, Δ_{sr} and Θ_{sr}:

Assumption 2. *Homogeneity assumption of* Γ_s, Δ_{sr} *and* Θ_{sr}

(H) *Homogeneity of the response probablities holds for* Γ_s, Δ_{sr} *and* Θ_{sr}*, hence,* $P(\mathbf{X}_{sr} = \mathbf{x}|\Theta_{sr}) = P(\mathbf{X}_{sr} = \mathbf{x}|\Gamma_s, \Delta_{sr})$

Let

$$P(X_{sri} \geq x|\Theta_{sr}) = \sum_{y=x}^{m} P(X_{sri} = y|\Theta_{sr}) \tag{5}$$

denote the probability of obtaining at least score x on item i given Θ_{sr}, which we refer to as the respondent-level item-step response function. For respondent r in group s, the expected item score is $E(X_{sri}|\Theta_r = \theta_r) = \sum_{x=1}^{m} P(X_{sri} \geq x|\Theta_{sr} = \theta_{sr})$. In two-level test data we distinguish between a respondent-level IRF (IRF-1, denoted $E_i(\cdot)$) and a group-level IRF (IRF-2, denoted $\mathcal{E}_i(\cdot)$). IRF-1 is defined as

$$\begin{aligned} E_i(\Theta_{sr}) &= E(X_{sri}|\Theta_{sr}) \\ &= \textstyle\sum_{x=1}^{m} P(X_{sri} \geq x|\Theta_{sr}), \end{aligned} \tag{6}$$

where $E(X_{sri}|\Theta_{sr})$ equals the expected item score Θ_{sr}.

Let $P(X_{sri} \geq x|\Gamma_s)$ denote the probability of obtaining at least score x on item i given Γ_s, which we refer to as the group-level item-step response function. By H and the law of total expectation (e.g., Rice, 2006, p 149), the item-step response function can be formulated as

$$\begin{aligned} P(X_{sri} \geq x|\Gamma_s) &= E[P(X_{sri} \geq x|\Theta_{sr}, \Gamma_s)|\Gamma_s] \\ &= E[P(X_{sri} \geq x|\Gamma_s + \Delta_s, \Gamma_s)|\Gamma_s] \\ &= E[P(X_{sri} \geq x|\Delta_s, \Gamma_s)|\Gamma_s] \\ &= E[P(X_{sri} \geq x|\Theta_{sr})|\Gamma_s]. \end{aligned} \tag{7}$$

For a randomly selected respondent in group s, the expected item score is

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

$E(X_{sri}|\Gamma_s = \gamma_s) = \sum_{x=1}^{m} P(X_{sri} \geq x|\Gamma_s = \gamma_s)$. IRF-2 is defined as

$$\begin{aligned} \mathcal{E}_i(\Gamma_s) &= E(X_{sri}|\Gamma_s) \\ &= \textstyle\sum_{x=1}^{m} P(X_{sri} \geq x|\Gamma_s) \\ &= E[\textstyle\sum_{x=1}^{m} P(X_{sri} \geq x|\Theta_{sr})|\Gamma_s] \quad \text{(Eq. 7)} \\ &= E[E_i(\Theta_{sr})|\Gamma_s] \quad \text{(Eq. 6)}, \end{aligned} \tag{8}$$

where $E(X_{sri}|\Gamma_s)$ equals the expected item score as a function of Γ_s. Note that because Δ_{sr} variables are assumed i.i.d., $E(X_{si}|\Gamma_s) = R^{-1}\sum_{r=1}^{R}\mathcal{E}_i(\Gamma_s) = \mathcal{E}_i(\Gamma_s)$. Hence, the expected group-level item score for group s is the value of the IRF-2 for $\Gamma_s = \gamma_s$. Figure 1 shows an hypothetical IRF-1 and IRF-2. Because IRF-2 is the expectation of IRF-1 with respect to Δ_{sr} (Equation 8), IRF-2 is flatter than function IRF-1.

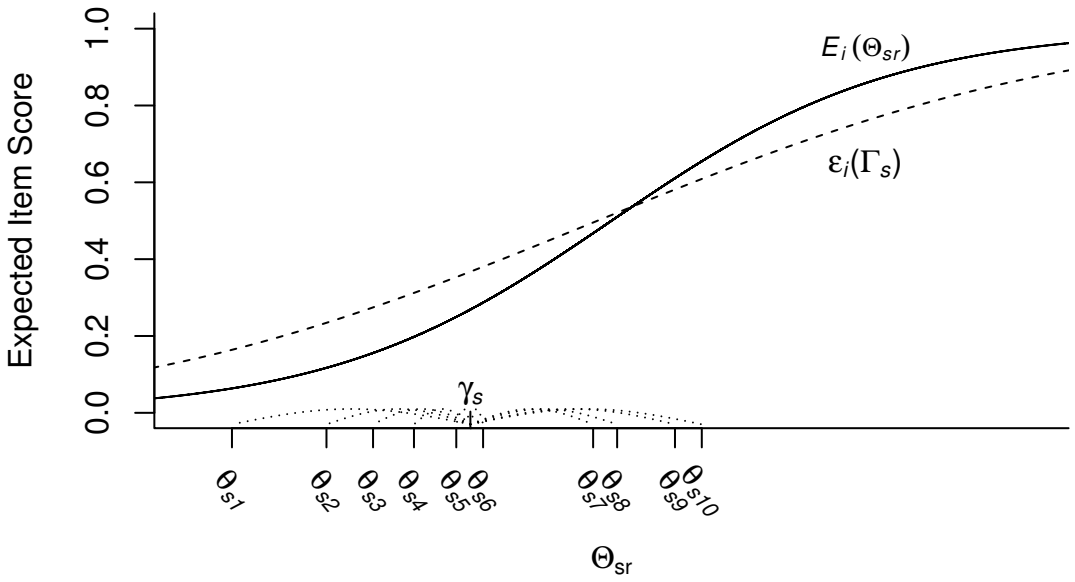

Figure 1: An IRF-1 ($E_i(\Theta_{sr})$; solid curve) and an IRF-2 ($\mathcal{E}_i(\Gamma_s)$; dashed curve), depicted on the same Θ_{sr} scale. The horizontal axis shows one hypothetical group value γ_s, plus the θ_{sr} values of 10 randomly drawn respondents ($r = 1, \ldots, 10$) from group s. Note that δ_{sr} is represented by the length of the line segment between γ_s and the θ_{sr} values on the horizontal axis.

2.1 Definitions of Possible Model Assumptions

Besides the basic and homogeneity assumption (B and H, respectively), multiple assumptions of nonparametric IRT for two-level data can be defined at level 1 (the respondent level) and at level 2 (the group level).

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

Definition 1. *Unidimensionality (UN).*

(UN-1) *Unidimensionality at level 1 holds if* Θ_{sr} *is a unidimensional variable.*

(UN-2) *Unidimensionality at level 2 holds if* Γ_s *is a unidimensional variable.*

UN-1 and UN-2 mean that the item scores on the test or questionnaire are modeled using one latent variable.

Definition 2. *Local independence (LI).*

(LI-1) *Local independence at level 1 holds if*

$$P(\mathbf{X}_{sr} = \mathbf{x}_{sr}|\Theta_{sr} = \theta_{sr}) \;=\; \prod_{i=1}^{I} P(X_{sri} = x_{sri}|\Theta_{sr} = \theta_{sr}) \tag{9}$$

(LI-2) *Local independence at level 2 holds if*

$$P(\mathbf{X}_{s} = \mathbf{x}_{s}|\Gamma_{s} = \gamma_{s}) = \prod_{r=1}^{R} P(\mathbf{X}_{sr} = \mathbf{x}_{sr}|\Gamma_{s} = \gamma_{s}) \tag{10}$$

LI-1 means that respondent-level item scores (X_{sri}) are independent given θ_{sr}. LI-2 means that the response vectors of respondents are independent given γ_s. LI-2 implies that between respondents, the respondent-level item scores X_{sri} and X_{spj} ($i \neq j$; $r \neq p$) are independent given γ_s. However, within respondents, respondent-level item scores X_{sri} and X_{srj} ($i \neq j$) are not independent given γ_s.

Definition 3. *Monotonicity (MO).*

(MO-1) *Monotonicity at level 1 holds if* $P(X_{sri} \geq x|\Theta_{sr} = \theta_{sr})$ *is nondecreasing in* θ_{sr}*, for all* i *and* $x = 1, \ldots, m$.

(MO-2) *Monotonicity at level 2 holds if* $P(X_{sri} \geq x|\Gamma_{s} = \gamma_{s})$ *is nondecreasing in* γ_s*, for all* i *and* $x = 1, \ldots, m$.

MO-1 implies that, for each item, IRF-1 (Equation 6) is nondecreasing in Θ_{sr}, and MO-2 implies that, for each item, IRF-2 (Equation 8) is nondecreasing in Γ_s. Note that in Figure 1, IRF-1 satisfies MO-1 and IRF-2 satisfies MO-2.

Definition 4. *Invariant item ordering (IIO). For a set of* I *items with* $m + 1$ *ordered item-score categories, for which the items are ordered and numbered such that* $E(X_{sri}) \leq E(X_{srj})$ *for all* $i < j$*, then*

(IIO-1) *Invariant item ordering at level 1 holds if* $E(X_{sri}|\Theta_{sr} = \theta_{sr}) \leq E(X_{srj}|\Theta_{sr} = \theta_{sr})$ *for all* θ_{sr}.

(IIO-2) *Invariant item ordering at level 2 holds if* $E(X_{sri}|\Gamma_{s} = \gamma_{s}) \leq E(X_{srj}|\Gamma_{s} = \gamma_{s})$ *for all* γ_s.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

IIO-1 means that the IRF-1s of different items do not intersect. IIO-2 means that the IRF-2s of different items do not intersect. Note that the definition of IIO-1 and IIO-2 allows for ties, such that for some values of the latent variable items may be equally difficult.

2.2 Relation Between Level 1 and Level 2 Assumptions

Theorem 1 gives the relations between the basic assumption and local independence at both levels.

Theorem 1. *B implies LI-1 and LI-2.*

The assumptions UN, LI, MO, and IIO were defined at both Level 1 and Level 2 (Definitions 1 to 4). However, Theorem 1 shows that B implies both LI-1 and LI-2, and as a result, LI is no longer a necessary assumption, as in all remaining proofs LI-1 and LI-2 may be replaced by B and H.

Theorem 2 gives the relations between the assumptions at level 1 and the assumptions at level 2.

Theorem 2. *Under B and H, UN-1, MO-1, and IIO-1 imply UN-2, MO-2, and IIO-2, respectively.*

Theorem 2 shows that the level-1 assumptions imply their level-2 assumptions, but not the other way around. Hence, the level-2 assumptions do not imply the level-1 assumptions. For example, if respondent-level item scores depend both on Γ_s and on Δ_{sr} and var(Δ_{sr})$>$ 0, in general $\Theta_{sr} \neq \Gamma_s$, $P(X_{sri} \geq x|\Theta_{sr}) \neq P(X_{sri} \geq x|\Gamma_s)$, and $E_i(\Theta_{sr}) \neq \mathcal{E}_i(\Gamma_s)$. As a result, UN-1, MO-1, and IIO-1 are not equal to UN-2, MO-2, and IIO-2, respectively. It may be noted that because of the homogeneity assumption H, the level-1 assumptions (UN-1, LI-1, MO-1, and IIO-1) are equivalent to the single-level nonparametric-IRT assumptions (UN, LI, MO and IIO), when X_{sri} is replaced by X_{ri} and θ_{sr} by θ_r.

2.3 Models

Two-level nonparametric IRT assumptions can be used to define several nonparametric IRT models. Analogous to the single-level nonparametric IRT models, we distinguish between the MHM and the DMM, but in addition we also distinguish between the level on which they can be defined. Snijders (2001) defined a two-level nonparametric IRT model for scaling groups with dichotomous item scores using assumptions UN-1, LI-1, MO-1, and IIO-1. We present four models

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

that allow for both dichotomous and polytomous items. As mentioned before, for all models B and H are assumed.

The first respondent-level model is the MHM-1, defined by assuming UN-1, LI-1, and MO-1 (Table 1, first row). The MHM-1 consists of level-1 assumptions, which imply UN-2, LI-2, and MO-2 (Theorem 2). The second respondent-level model is the DMM-1, defined by assuming UN-1, LI-1, MO-1, and IIO-1, implying UN-2, LI-2, MO-2, and IIO-2 (Table 1, second row). The first group-level model is the MHM-2, defined by assuming UN-2, LI-2, and MO-2 (Table 1, third row). The second group-level model is the DMM-2, defined by assuming UN-2, LI-2, MO-2, and IIO-2 (Table 1, fourth row). Note that, for all models, LI-1 and LI-2 are implied by B, but we explicitly incorporate them into the models, such that the models align more obviously to the single-level models.

Table 1:

Assumptions of the Two-Level Nonparametric IRT Models.

Model	Respondent-level assumptions				Group-level assumptions			
	UN-1	LI-1	MO-1	IIO-1	UN-2	LI-2	MO-2	IIO-2
MHM-1	A	A	A		I	I	I	
DMM-1	A	A	A	A	I	I	I	I
MHM-2					A	A	A	
DMM-2					A	A	A	A

Note. A = assumed, I = implied.

Figure 2 shows the hierarchical structure of the four models, where an arrow indicates an implication. The MHM-2 is the most general model, of which the other three are special cases. The DMM-1 is the most restrictive model, implying the other three models. In the next sections we derive some ordering and observable properties implied by these models.

3 Ordering Properties of Two-Level Nonparametric IRT Models

We investigated four possible ordering properties for sum score X_{sr+} at level 1, and sum score X_{s+} at level 2: MLR, SOM, SOL, and weak SOL.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

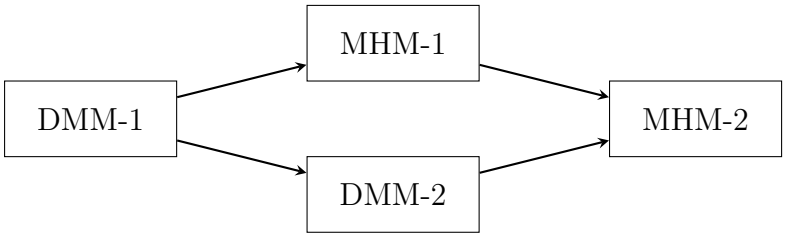

Figure 2: Hierarchical Structure of the Two-Level Nonparametric IRT Models.

Definition 5. *Monotone likelihood ratio (MLR; Ferguson, 1967, p. 208).*

(MLR-1) *Monotone likelihood ratio at level 1 holds if, for $a < b$, the probability ratio*

$$\frac{P(X_{sr+} = b|\Theta_{sr})}{P(X_{sr+} = a|\Theta_{sr})} \text{ is nondecreasing in } \Theta_{sr}. \tag{11}$$

(MLR-2) *Monotone likelihood ratio at level 2 holds if, for $a < b$, the probability ratio*

$$\frac{P(X_{s+} = b|\Gamma_s)}{P(X_{s+} = a|\Gamma_s)} \text{ is nondecreasing in } \Gamma_s. \tag{12}$$

Definition 6. *Stochastic ordering of the manifest score by the latent variable (SOM; Hemker et al., 1997).*

(SOM-1) *Stochastic ordering of the manifest score by the latent variable at level 1 holds if, for any value x and $t < u$*

$$P(X_{sr+} \geq x|\Theta_{sr} = t) \leq P(X_{sr+} \geq x|\Theta_{sr} = u). \tag{13}$$

(SOM-2) *Stochastic ordering of the manifest score by the latent variable at level 2 holds if, for any value x and $t < u$*

$$P(X_{s+} \geq x|\Gamma_s = t) \leq P(X_{s+} \geq x|\Gamma_s = u). \tag{14}$$

Definition 7. *Stochastic ordering of the latent variable by the manifest score (SOL; Hemker et al., 1997).*

(SOL-1) *Stochastic ordering of the latent variable by the manifest score at level 1 holds if, for any value t and $a < b$,*

$$P(\Theta_{sr} > t|X_{sr+} = a) \leq P(\Theta_{sr} > t|X_{sr+} = b). \tag{15}$$

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

(SOL-2) *Stochastic ordering of the latent variable by the manifest score at level 2 holds if, for any value t and $a < b$,*

$$P(\Gamma_s > t|X_{s+} = a) \leq P(\Gamma_s > t|X_{s+} = b). \tag{16}$$

Definition 8. *Weak SOL (WSOL; Van der Ark & Bergsma, 2010).*

(WSOL-1) *Weak SOL at level 1 holds if, for any t and a*

$$P(\Theta_{sr} > t|X_{sr+} < a) \leq P(\Theta_{sr} > t|X_{sr+} \geq a). \tag{17}$$

(WSOL-2) *Weak SOL at level 2 holds if, for any t and a*

$$P(\Gamma_s > t|X_{s+} < a) \leq P(\Gamma_s > t|X_{s+} \geq a). \tag{18}$$

In general, ordering property MLR implies SOM, SOL, and WSOL, and SOL implies WSOL (Hemker et al., 1997; Lehmann, 1986, p. 85; Van der Ark & Bergsma, 2010). Hence, ordering property MLR-1 implies SOM-1, SOL-1, and WSOL-1, whereas ordering property MLR-2 implies SOM-2, SOL-2, and WSOL-2. The MLR, SOM, SOL, and WSOL results are valid for any monotone nondecreasing item summary within respondents (e.g., all-correct score, rest-scores, subscores; Rosenbaum, 1984).

For two-level test data, it is unknown whether MLR, SOM, SOL, or WSOL are implied by the two-level nonparametric IRT models. Theorem 3 gives the result for the strongest ordering property (MLR) for the least restrictive models (MHM-1 and MHM-2) and generalizes to weaker ordering properties and more restrictive models.

Theorem 3.

(a) For dichotomous item scores, the MHM-1 implies MLR-1.

(b) For dichotomous item scores, for $R \geq I$, the MHM-2 implies MLR-2.

MLR is symmetric in its argument, so the statement X_{sr+} has MLR in Θ_{sr} means that Θ_{sr} also has MLR in X_{sr+}. Theorem 3 implies that for dichotomous items, under the MHM-1 X_{sr+} is stochastically ordered by Θ_{sr} (SOM-1) and Θ_{sr} is stochastically ordered by X_{sr+} (SOL-1). It may be noted that Theorem 3(a) is very similar to the result obtained by Grayson (1988) who proved for single-level dichotomous item scores that the MHM implies MLR. Under the MHM-2, for $R \geq I$, group-level item score X_{s+} is stochastically ordered by Γ_s (SOM-2) and Γ_s is stochastically ordered by X_{s+} (SOL-2). Note that for $R < I$, MLR-2 is implied for the sum score of any random subset of items of size I^*, for which $I^* \leq R$. Because the DMM-1 is a special case of the MHM-1 (see Figure 2), Theorem 3(a)

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

also applies to the DMM-1. Similarly, the MHM-1, the DMM-1, and the DMM-2 are special cases of the MHM-2 (see Figure 2), Theorem 3(b) applies to these models as well.

For polytomous items, the single-level MHM and DMM generally do not imply MLR and SOL (see Hemker et al., 2001, for counter examples) but these models do imply SOM (Hemker et al., 1996, 1997) and weak SOL (Van der Ark & Bergsma, 2010). Theorem 4 and 5 show that these results generalize to two-level models.

Theorem 4.

(a) The MHM-1 implies SOM-1.

(b) The MHM-2 implies SOM-2.

Theorem 4 implies that under the MHM-1 X_{sr+} is stochastically ordered by Θ_{sr} (SOM-1) and under the MHM-2, X_{s+} is stochastically ordered by Γ_s (SOM-2). Because the DMM-1 is a special case of the MHM-1, it also implies SOM-1. Also, because the MHM-1, the DMM-1, and the DMM-2 are special cases of MHM-2, these models imply SOM-2.

Theorem 5.

(a) The MHM-1 implies WSOL-1.

(b) The MHM-2 implies WSOL-2.

Let $\mathbf{1}(X_{sr+} \geq k)$ denote the dichotomized respondent-level sum score that takes on value 1 if $X_{sr+} \geq k$, and 0 otherwise, and let $\mathbf{1}(X_{s+} \geq k)$ denote the dichotomized group-level sum score that takes on value 1 if $X_{s+} \geq k$, and 0 otherwise. Then, Theorem 5 implies that under the MHM-1, Θ_{sr} is stochastically ordered by $\mathbf{1}(X_{sr+} \geq k)$ (WSOL-1), and that under the MHM-2, Γ_s is stochastically ordered by $\mathbf{1}(X_{s+} \geq k)$ (WSOL-2). Because DMM-1 is a special case of MHM-1, this model also implies WSOL-1. Also, because the MHM-1, the DMM-1, and the DMM-2 are special cases of the MHM-2, these models imply WSOL-2.

4 Observable Properties of Two-Level Nonparametric IRT Models

We define observable properties CA, MM, and MIIO for two-level IRT models. For single-level IRT models, rest score $X_{r(i)}$ was used in MM, and rest score $X_{r(ij)}$ was used in NNIIC and in MIIO. These rest scores are proxies for the latent variable

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

that must be independent of the variables under investigation. Because of the independence requirement, for two-level IRT models, these rest scores become more involved. Table 2 provides an overview of these rest scores for classified by observable property and level. The rest scores at Level 1 can be considered within-respondent rest scores, the rest scores at Level 2 can be considered between-level rest scores.

Table 2:

Overview of rest scores used in observable properties in single-level and two-level IRT models.

Observable Property	Single-Level IRT	Two-Level IRT	
		Level 1	Level 2
MM	$X_{r(i)} = \sum_{j\neq i}^{I} X_{rj}$	$X_{sr(i)} = \sum_{j\neq i}^{I} X_{srj}$	$X_{r(r,i)} = \frac{\sum_{p\neq r}^{R}\sum_{j\neq i}^{I} X_{spj}}{R-1}$
NNIIC[a]	$X_{r(ij)} = \sum_{k\neq i,j}^{I} X_{rk}$	$X_{sr(ij)} = \sum_{k\neq i,j}^{I} X_{srk}$	$X_{r(rp,ij)} = \frac{\sum_{q\neq r,p}^{R}\sum_{k\neq i,j}^{I} X_{sqk}}{R-2}$
MIIO	$X_{r(ij)} = \sum_{k\neq i,j}^{I} X_{rk}$	$X_{sr(ij)} = \sum_{k\neq i,j}^{I} X_{srk}$	$X_{r(r,ij)} = \frac{\sum_{p\neq r}^{R}\sum_{k\neq i,j}^{I} X_{spk}}{R-1}$

a: Pertains to the NNIIC given the rest score. The other two NNIIC inequalities do not use a rest score.

Definition 9 defines CA for two-level IRT models. First, partition $\mathbf{X}_{sr}$ into two mutually exclusive and exhaustive sets $\mathbf{Y}_{sr}$ and $\mathbf{Z}_{sr}$. For example, $\mathbf{Y}_{sr}$ may contain X_{sr1} and X_{sr2} and $\mathbf{Z}_{sr}$ the remaining item scores. Second, partition the response vectors of the R respondents in group s — $\mathbf{X}_{s1}, \ldots, \mathbf{X}_{sR}$ — which are collected in $\mathbf{X}_s$, into three mutually exclusive and exhaustive sets: $\mathbf{Y}_{s1}$, $\mathbf{Y}_{s2}$, and $\mathbf{Z}_s$. For example, $\mathbf{Y}_{s1}$ could contain just $\mathbf{X}_{sr}$, $\mathbf{Y}_{s1}$ could contain just $\mathbf{X}_{s2}$, and $\mathbf{Z}_s$ could contain the remaining response vectors from $\mathbf{X}_s$. Note that all scores of the same respondent are in the same set.

Definition 9. *Conditional association (CA; Holland & Rosenbaum, 1986; Rosenbaum, 1988).*

(CA-1) Conditional association at level 1 holds if

$$Cov[g_1(\mathbf{Y}_{sr}), g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] \geq 0. \tag{19}$$

(CA-2) Conditional association at level 2 holds if, for $r \neq p$,

$$Cov[g_1(\mathbf{Y}_{sr}), g_2(\mathbf{Y}_{sp})|h(\mathbf{Z}_{s(rp)})] \geq 0. \tag{20}$$

CA-1 is conditional association of the scores within respondents, whereas CA-2 is conditional association of the scores between respondents in the same group (see, also, Rosenbaum, 1988). As for CA, CA-1 and CA-2 are too comprehensive

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

for a single test procedure. The testing procedure to identify locally dependent item sets using NNIIC (Straat et al., 2016) can be readily generalized to two-level models: For Level 1, the three inequalities in NNIIC generalize to $Cov(X_{sri}, X_{srj}) \geq 0$, $Cov(X_{sri}, X_{srj}|X_{srk}) \geq 0$, and $Cov(X_{sri}, X_{srj}|X_{sr(ij)}) \geq 0$. For Level 2, let q, q, and r index three different respondents. The three inequalities generalize to $Cov(X_{sri}, X_{spj}) \geq 0$, $Cov(X_{sri}, X_{spj}|X_{sqk}) \geq 0$, and $Cov(X_{sri}, X_{spj}|X_{s(rp,ij)}) \geq 0$. Rest scores $X_{sr(ij)}$ and $X_{s(rp,ij)}$ have been defined in Table 2.

Definition 10 defines MM for two-level IRT models. The rest scores used in Definition 10 have been defined in Table 2

Definition 10. *Manifest monotonicity (MM; Junker, 1993; Sijtsma & Hemker, 2000).*

(MM-1) *Manifest monotonicity at level 1 holds if the within-respondent item-rest regression $E(X_{sri}|X_{sr(i)})$ is nondecreasing in $X_{sr(i)}$.*

(MM-2) *Manifest monotonicity at level 2 holds if the between-respondent item-rest regression $E(X_{sri}|X_{s(r,i)})$ is nondecreasing in $X_{s(r,i)}$.*

Definition 11 defines MM for two-level IRT models. The rest scores used in Definition 11 have been defined in Table 2

Definition 11. *Manifest invariant item ordering (MIIO; Ligtvoet et al., 2010).*

(MIIO-1) *Manifest invariant item ordering at level 1 holds if, for $E(X_{sri}) < E(X_{srj})$, $E(X_{sri}|X_{sr(ij)} = y) \leq E(X_{srj}|X_{sr(ij)} = y)$ for all y and all $i < j$.*

(MIIO-2) *Manifest invariant item ordering at level 2 holds if, for $E(X_{sri}) < E(X_{srj})$, $E(X_{sri}|X_{s(r,ij)} = y) \leq E(X_{srj}|X_{s(r,ij)} = y)$ for all y and all $i < j$.*

In Theorem 7, 6, and 8 we state which two-level models imply the observable properties CA, MM, and MIIO, respectively.

Theorem 6.

(a) The MHM-1 implies CA-1.

(b) The MHM-2 implies CA-2.

Because the DMM-1 is a special case of the MHM-1, it also implies CA-1. Also, because the MHM-1, the DMM-1, and the DMM-2 are special cases of MHM-2, these models imply CA-2.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

Theorem 7.

(a) For dichotomous items, the MHM-1 implies MM-1.

(b) For dichotomous items, the MHM-2 implies MM-2.

Because the DMM-1 is a special case of the MHM-1, it also implies MM-1. Also, because the MHM-1, the DMM-1, and the DMM-2 are special cases of MHM-2, these models imply MM-2. As for single-level IRT models, MM does not necessarily hold for polytomous items. However, MM-1 and MM-2 may still provide heuristic evidence for or against the MHM-1 and/or the MHM-2 (cf., Sijtsma & Van der Ark, 2020, p. 151). Alternatively, if polytomous items are dichotomized, Theorem 7 holds (Junker & Sijtsma, 2000).

Theorem 8.

(a) The DMM-1 implies MIIO-1.

(b) The DMM-2 implies MIIO-2.

Because the DMM-1 is a special case of the DMM-2, it also implies CA-2.

5 Relations Between Models and Properties

In the previous sections we defined four models, eight ordering properties, and six observable properties. In addition, we provided proofs for which model implied which property, for the least restrictive model and strongest property possible. Because more restrictive models are special cases of models with fewer restrictions, they are defined with at least the same assumptions that imply the property (see Figure 2). Hence, more restrictive models imply the same properties as the more general models.

Table 3 provides an overview of the most important implications for each model. The MHM-1 (Table 3, first column) implies (W)SOL-1 and (W)SOL-2. Hence, the MHM-1 implies an ordinal respondent-level scale, on which respondents may be stochastically ordered on Θ_{sr} using X_{sr+}, and an ordinal group-level scale, on which groups may be stochastically ordered on Γ_s using X_{s+}. Methods for investigating the model fit of the MHM-1 are MM-1, CA-1, MM-2, and CA-2. In addition to the implications by the MHM-1, the DMM-1 (Table 3, second column) also implies an ordinal item scale on which items may be stochastically ordered on their latent difficulty using the mean scores on the items. Methods MIIO-1 and MIIO-2 can be

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

Table 3:

Implied Properties of the Two-Level Nonparametric IRT Models.

Ordering Property	Model			
	MHM-1	DMM-1	MHM-2	DMM-2
MLR-1	D	D		
SOL-1	D	D		
SOM-1	A	A		
WSOL-1	A	A		
MLR-2	D	D	D	D
SOL-2	D	D	D	D
SOM-2	A	A	A	A
WSOL-2	A	A	A	A
Observable Property	Model			
	MHM-1	DMM-1	MHM-2	DMM-2
MM-1	D	D		
CA-1	A	A		
MIIO-1		A		
MM-2	D	D	D	D
CA-2	A	A	A	A
MIIO-2		A		A

Note. A = property is implied for dichotomous and polytomous items, D = property is implied for dichotomous or dichotomized items only.

used for investigating model fit of the DMM-1 in addition to the methods of the MHM-1.

The MHM-2 (Table 3, third column) implies (W)SOL-2. Hence, the MHM-2 implies an ordinal group-level scale on which groups may be stochastically ordered on Γ_s using X_{s+}. Methods for investigating the model fit of the MHM-2 are MM-2 and CA-2. In addition to the implications by the MHM-2, the DMM-2 (Table 3, fourth column) also implies an ordinal item scale on which items may be stochastically ordered on their latent difficulty using the mean scores on the items. Methods MIIO-1 and MIIO-2 can be used for investigating model fit of the DMM-2 in addition to the methods of the MHM-2.

The two-level nonparametric IRT models are defined on either or both the respondent level and the group level. Depending on the interest of the researcher, one or both levels are relevant for scaling. If the goal is to scale the respondents, it is

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

sufficient to mainly focus on checking the respondent-level assumptions of the MHM-1 or DMM-1. If the goal is to only scale the groups, as is the case in multi-rater data, the group-level assumptions are of key interest. For example, if a group-level IRF is flat, an item does not discriminate between low and high values of Γ_s. Such an item does not contribute to accurate measurement on the group level. In addition, the respondent-level assumptions are informative for investigating, for example, whether the respondents may also be ordered using their sum score, or how the results relate to each other across levels. Therefore, even though investigating the MHM-2 or DMM-2 is sufficient to determine model fit at the group level, investigating the MHM-1 or the DMM-1 by checking assumptions on both level 1 and level 2 is suggested. If model violations occur at level 1, it is still possible that there are no violations at level 2, and the MHM-2 or the DMM-2 fit the data.

6 Discussion

The main contribution of this paper is the establishment of ordering properties and observable properties for two-level nonparametric IRT models. Ordering properties MLR-1, MLR-2, SOL-1.SOL-2, weak SOL-1, weak-SOL-2 SOM-1, and SOM-2 justify ordinal measurement using two-level nonparametric IRT models, in a way that is similar to ordinal measurement in the more popular single-level nonparametric IRT models. In addition, the observable properties MM-1, MM-2, CA-1, CA-2, MIIO-1, and MIIO-2 allow researchers to investigate the fit of the two-level nonparametric IRT models". Combined, these newly established ordering properties and observable properties enables the practical use of the two-level measurement models

Building on previous work by Snijders (2001), we introduced four models for two-level test data. For level 1, we introduced the MHM-1, which allows ordering nested respondents on latent variable Θ_{sr} using manifest variable X_{sr+}, and the DMM-1, which allows ordering nested respondents and items on Θ_{sr} using X_{sr+} and $E(X_{sri})$, respectively. For level 2, we introduced the MHM-2, which allows ordering groups on latent variable Γ_s using manifest variable X_{s+}, and the DMM-2, which allows ordering groups and items on Γ_s using X_{s+} and $E(X_{si})$, respectively. The hierarchical relations among the four models shows that the DMM-1 implies all other models and that the MHM-2 is the most general model (see Figure 2).

In addition, we derived observable data properties implied by the models, which can be used to investigate the model fit for a given data set. Specifically, we generalized the properties manifest monotonicity, conditional association, and

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

manifest invariant item ordering for the respondent level and the group level. Theorem 7(b) showed the perhaps surprising result that, for a test consisting of dichotomous items, even though group-level item scores are not dichotomous (because they combine the item scores across respondents), still the strong results for dichotomous nonparametric IRT models hold. In deriving level-2 properties from level-1 properties, assuming the individual respondent-variables Δ_{sr} are i.i.d. proved to be a key ingredient. Assuming i.i.d. in test data is usually based on the sampling design or data collection conditions in relation to the latent variable. However, finding support for the i.i.d. assumption based on observable properties on the group level may be a valuable topic for future research.

The properties derived in this paper apply at the population level. Koopman et al. (2023) suggested statistical tests for MO-1, MO-2, IIO-1, and IIO-2 using observable properties MM-1, MM-2, MIIO-1, and MIIO-2, respectively. Using simulated data, these authors found that the tests for MO-1, IIO-1 and IIO-2 had satisfactory Type-1 error rates and power, whereas the tests for MO-2 had satisfactory Type-1 error-rates but insufficient power (see also, Koopman, 2023). Note that both procedures deviated slightly from the results in this paper, because they used level-2 item scores rather than the between-respondent item scores that were used in the MM-2 and MIIO-2 definitions in this paper. Perhaps these latter item scores increases the power of the significance test of MO-2.

Note that Molenaar (1997) originally defined the DMM non-intersecting item-step response functions $P(X_i \geq x|\Theta)$ rather than an IRT model having non-intersecting item-response functions. As investigating properties of items can be considered more relevant than investigating properties of item-steps, the new definition of the DMM in terms of non-intersecting IRFs can be considered more useful. In addition, the property of IIO is defined in terms of conditional expected item scores, and fits better to the new definition of the DMM than to the original definition. If there is reason to require an invariant item-step order, an alternative DMM-like model may be proposed including this assumption. However, one should realize that an invariant item-step order not necessarily implies an invariant item order (Sijtsma & Hemker, 1998).

In this paper we chose to expand on work by Snijders (2001), because of its strong link to the one-level MHM and DMM. However, other generalizations of the MHM and DMM are possible. Within the framework of this paper, one may also consider a within-group model, in which the IRFs are assumed to be increasing only in δ_{sr}. Such a model may be useful if the focus is on within-group comparison only rather than comparison of all respondents, or if items contain a relative component in relation to

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

a group aspect. Properties and applications of this model are yet unknown. Outside the framework proposed in this paper, Koopman, Zijlstra, De Rooij, and Van der Ark (2020) proposed the nonparametric hierarchical rater model, a nonparametric version of the (parametric) hierarchical rater model (Patz et al., 2002). Possibly other two-level parametric IRT models may be redefined as a nonparametric model, such as the multiple raters model (Verhelst & Verstralen, 2001) or the rater bundle model (Wilson & Hoskens, 2001). Alternatively, the nonparametric partial credit model or nonparametric sequential model (Hemker et al., 1997, 2001, respectively) may be generalized to a two-level framework.

The presented models in this paper are unidimensional models. Hence, for MHM-1 and DMM-1, it is assumed that respondents across groups may be located on the same latent variable. This is quite a strict assumption and whether this is sensible should be investigated, for example by analysis on differential item functioning (Holland & Wainer, 1993). Known methods within nonparametric IRT are comparing scales and scale properties across groups (Sijtsma & Van der Ark, 2017; Van der Ark et al., 2008) and performing an IIO analysis (Sijtsma & Junker, 1996). Two-level IRT modeling may benefit from multidimensional generalizations for developing scales that explicitly separate a respondent and group dimension. How these alternative models hierarchically relate to the models presented in this paper, and what properties they imply, is a topic for further investigation.

The developments presented in this paper are part of a larger project to make all elements of Mokken scale analysis available for two-level test data (Koopman, Zijlstra, & Van der Ark, 2020; Koopman et al., 2022). Next steps in development should be aimed at developing group-level item selection procedures and at allowing more complex research designs, such as a cross nested design in which respondents score multiple groups.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

References

Ahmed, A.-H. N., Leon, R., & Proschan, F. (1981). Generalized association, with applications in multivariate statistics. *The Annals of Statistics*, *9*(1), 168–176. http://doi.org/10.1214/aos/1176345343

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick. *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

Cann, A., Watson, A. J., & Bridgewater, E. A. (2014). Assessing humor at work: The humor climate questionnaire. *Humor: International Journal of Humor Research*, *27*(2). http://doi.org/10.1515/humor-2014-001

Chang, H.-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika*, *59*(3), 391–404. http://doi.org/10.1007/BF02296132

De Jong, M. G., & Steenkamp, J.-B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika*, *75*, 3–32. http://doi.org/10.1007/S11336-009-9134-Z

Dougherty, M. B., & Larson, E. L. (2010). The nurse-nurse collaboration scale. *The Journal of Nursing Administration*, *40*(1), 17–25. http://doi.org/10.1097/NNA.0b013e3181c47cd6

Edelsbrunner, P. (2022). A model and its fit lie in the eye of the beholder: Long live the sum score. *Frontiers in Psychology*, *13*, 1–5. http://doi.org/10.3389/fpsyg.2022.986767

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika*, *79*(2), 303–316. http://doi.org/10.1007/s11336-013-9341-5

Ellis, J. L., & Sijtsma, K. (2023). A test to distinguish monotone homogeneity from monotone multifactor models. *psychometrika*, *88*(2), 387–412. http://doi.org/10.1007/s11336-023-09905-w

Esary, J. D., Proschan, F., & Walkup, D. W. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics*, *38*(5), 1466–1474. http://doi.org/10.1214/aoms/1177698701

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach.* Academic Press.

Fox, J.-P. (2007). Multilevel IRT modeling in practice with the package mlirt.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

Journal of Statistical Software, *20*, 1–16. http://doi.org/10.18637/jss.v020.i05

Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271–288. http://doi.org/10.1007/BF02294839

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*(3), 383–392. http://doi.org/10.1007/BF02294219

Hemker, B. T., Sijtsma, K., Molenaar, I., & Junker, B. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*(4), 679–693. http://doi.org/10.1007/BF02294042

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*(3), 331–347. http://doi.org/10.1007/BF02294555

Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*(4), 487–506. http://doi.org/10.1007/BF02296191

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*(4), 1523–1543. http://doi.org/10.1214/aos/1176350174

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Routledge.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent bernoulli random variables. *Psychometrika*, *59*(1), 77–79. http://doi.org/10.1007/BF02294266

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, *56*(2), 255–278. http://doi.org/10.1007/BF02294462

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, *21*(3), 1359–1378. http://doi.org/10.1214/aos/1176349262

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics*, *25*(3), 1327–1343. http://doi.org/10.1214/aos/1069362751

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*(1), 65–81. http://doi.org/10.1177/01466216000241004

Kamae, T., Krengel, U., & O'Brien, G. L. (1977). Stochastic inequalities on partially ordered spaces. *The Annals of Probability*, *5*(6), 899–912. http://doi.org/

10.1214/aop/1176995659

Koopman, L. (2023). Effect of within-group dependency on fit statistics in Mokken scale analysis in the presence of two-level test data. In M. Wiberg, D. Molenaar, J. González, & J.-S. Kim (Eds.), *Quantitative psychology: The 87th annual meeting of the Psychometric Society, Bologna, Italy, 2022.* Springer.

Koopman, L., Zijlstra, B. J. H., De Rooij, M., & Van der Ark, L. A. (2020). Bias of two-level scalability coefficients and their standard errors. *Applied Psychological Measurement*, *44*(3), 197–214. http://doi.org/10.1177/0146621619843821

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2020). Standard errors of two-level scalability coefficients. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 213–236. http://doi.org/10.1111/bmsp.12174

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2022). A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. *Quality of Life Research*, *31*(1), 25–36. http://doi.org/10.1007/s11136-021-02840-2

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2023). Evaluating model fit in two-level mokken scale analysis. *Psych*, *5*(3), 847–865. http://doi.org/10.3390/psych5030056

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). Wiley.

Ligtvoet, R. (2022). Incomplete tests of conditional association for the assessment of model assumptions. *psychometrika*, *87*(4), 1214–1237. http://doi.org/10.1007/s11336-022-09841-1

Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika*, *76*(2), 200–216. http://doi.org/10.1007/S11336-010-9199-8

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. http://doi.org/10.1177/0013164409355697

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

Mokken, R. J. (1969). Dutch-American comparisons of the 'sense of political efficacy': some remarks on cross-cultural 'robustness'. *Acta Politica*, *4*(4), 425–448.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Mouton.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In

W. J. van der Linden & K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). Springer.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384. http://doi.org/10.3102/10769986027004341

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611–630. http://doi.org/10.1007/BF02294494

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Rice, J. A. (2006). *Mathematical statistics and data analysis* (3rd ed.). Thomson Brooks/Cole.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*(3), 425–435. http://doi.org/10.1007/BF02306030

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349–359. http://doi.org/10.1007/BF02294217

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometrika monograph supplement No. 17). Psychometric Society. http://www.psychometrika.org/journal/online/MN17.pdf

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in psychology*, *7*, 1–16. http://doi.org/10.3389/fpsyg.2016.00110

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders.* Springer.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*(2), 183–200. http://doi.org/10.1007/BF02294774

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, *25*(4), 391–415. http://doi.org/10.3102/10769986025004391

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*(1), 79–105. http://doi.org/10.1111/j.2044-8317.1996.tb01076.x

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191–207. http://doi.org/10.1007/

BF02294835

Sijtsma, K., Meijer, R. R., & Van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*(1), 31–37. http://doi.org/10.1016/j.paid.2010.08.016

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.

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. http://doi.org/10.1111/bmsp.12078

Sijtsma, K., & Van der Ark, L. A. (2020). *Measurement models for psychological attributes*. Chapman and Hall/CRC Press.

Snijders, T. A. B. (2001). Two-level non-parametric scaling for dichotomous data. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 319–338). Springer. http://doi.org/10.1007/978-1-4613-0169-1_17

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293–325. http://doi.org/10.1007/BF02295289

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology*, *12*(4), 117–123. http://doi.org/10.1027/1614-2241/a000115

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. http://doi.org/10.1007/S11336-011-9201-0

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55. http://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 179–187. http://doi.org/10.1348/000711007X173391

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*(3), 273–282. https://doi.org/10.1177/01466210122032073

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

Van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika*, *75*(2), 272–279. http://doi.org/10.1007/s11336-010-9147-7

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika*, *73*(2), 183–208. http://doi.org/10.1007/s11336-007-9034-z

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). Springer. http://doi.org/10.1007/978-1-4613-0169-1_5

Widaman, K. F., & Revelle, W. (2022). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, 1–19. http://doi.org/10.3758/s13428-022-01849-w

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306. http://doi.org/10.3102/10769986026003283

Appendix

Lemma A1. *UN-1 implies UN-2.*

Proof. Equation 4 defines $\Theta_{sr} = \Gamma_s + \Delta_{sr}$, which implies $\Gamma_s = \Theta_{sr} - \Delta_{sr}$. However, as Γ_s does not depend on r, $\Gamma_s = E(\Theta_{sr} - \Delta_{sr}) = E(\Theta_{sr}) - E(\Delta_{sr})$, where $E(\Theta_{sr})$ denotes the expectation of Θ_{sr} over the respondents in randomly selected s. Because $E(\Delta_{sr}) = 0$, it follows that $\Gamma_s = E(\Theta_{sr})$ within group s. If Θ_{sr} is unidimensional, its expectation $E(\Theta_{sr})$ is also unidimensional. Variable Θ_{sr} is unidimensional by UN-1; hence, variable Γ_s is also unidimensional. □

Lemma A2. *MO-1 implies MO-2.*

Proof. Let $P(\Delta_{sr})$ denote the probability density function of the distribution of Δ_{sr}. By H, the group-level item-step response function is,

$$\begin{aligned} P(X_{sri} \geq x|\Gamma_s) &= E[P(X_{sri} \geq x|\Theta_{sr})|\Gamma_s] \text{ (Equation 7)} \\ &= \int P(X_{sri} \geq x|\Theta_{sr})P(\Delta_{sr}|\Gamma_s)d\Delta_{sr} \end{aligned} \tag{A1}$$

As Δ_{sr} and Γ_s are independent by B, $P(\Delta_{sr}|\Gamma_s) = P(\Delta_{sr})$, and the last term of Equation A1 reduces to

$$\int P(X_{sri} \geq x|\Theta_{sr})P(\Delta_{sr})d\Delta_{sr} \tag{A2}$$

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

By MO-1, $P(X_{sri} \geq x|\Theta_{sr} = \theta_{sr})$ is nondecreasing in θ_{sr}. Hence, Equation A1 is nondecreasing in γ_s, which equals the definition of MO-2. □

Lemma A3. *IIO-1 implies IIO-2.*

Proof. By IIO-1

$$\begin{array}{lll} & E(X_{sri}|\Theta_{sr} = \theta_{sr}) & \leq E(X_{srj}|\Theta_{sr} = \theta_{sr}) \text{ for all } \theta_{sr} \\ \Leftrightarrow & E_i(\Theta_{sr}) & \leq E_j(\Theta_{sr}) \\ \Leftrightarrow & \int E_i(\Theta_{sr})P(\Delta_{sr}|\Gamma_s)d\Delta_{sr} & \leq \int E_j(\Theta_{sr})P(\Delta_{sr}|\Gamma_s)d\Delta_{sr} \\ \Leftrightarrow & E[E_i(\Theta_{sr})|\Gamma_s] & \leq E[E_j(\Theta_{sr})|\Gamma_s] \\ \Leftrightarrow & \mathcal{E}_i(\Gamma_s) & \leq \mathcal{E}_j(\Gamma_s) \text{ (by H, Eq. 8)} \\ \Leftrightarrow & E(X_{sri}|\Gamma_s = \gamma_s) & \leq E(X_{srj}|\Gamma_s = \gamma_s) \text{ for all } \gamma_s \end{array} \tag{A3}$$

The final result in Equation A3 equals the definition of IIO-2. □

Lemma A4. *MHM-1 implies that $E(g(\mathbf{X}_{sr})|\Theta_{sr} = \theta_{sr})$ is nondecreasing in θ_{sr} for any bounded, nondecreasing function $g(\cdot)$.*

Proof. By LI-1, scores X_{sri} within $\mathbf{X}_{sr}$ are independent given Θ_{sr}. By MO-1, X_{sri} is stochastically ordered in Θ_{sr}; that is, for $t < u$, $P(X_{sri} \geq x|\Theta_{sr} = t) \leq P(X_{sri} \geq x|\Theta_{sr} = u)$ for all i and all x. For a set of independent variables the stochastic ordering is preserved under convolutions, for any bounded, nondecreasing function $g(\cdot)$ (Shaked and Shanthikumar e.g., 2007, Theorem 1.A.3(b); see also Ahmed et al. 1981, Lemma 3.3; Holland and Rosenbaum 1986, Lemma 2). Hence

$$E[g(\mathbf{X}_{sr})|\Theta_{sr} = t] \leq E[g(\mathbf{X}_{sr})|\Theta_{sr} = u]. \tag{A4}$$

□

Lemma A5. *MHM-2 implies $E(g(\mathbf{X}_s)|\Gamma_s = \gamma_s)$ is nondecreasing in γ_s for any bounded, nondecreasing function $g(\cdot)$.*

Proof. Assumptions UN-2, LI-2, and MO-2 are equivalent to Rosenbaum's (1988) assumptions (1), (6), and (7), respectively, which collectively define the item-bundel model. In his Lemma 1, Rosenbaum showed that for any bounded, nondecreasing function $g(\cdot)$ for which UN-2, LI-2, and MO-2 holds, $E(g(\mathbf{X}_s)|\Gamma_s = \gamma_s)$ is nondecreasing in γ_s (see, also, Kamae et al., 1977, Proposition 1). □

Proof of Theorem 1. *(B implies LI-1 and LI-2)*

Proof. The independence of the Γ_s, Δ_{sr}, and ε_{sr} by B implies that the ε_{sr} are independent given $\Gamma_s + \Delta_{sr} = \Theta_{sr}$, and that the $(\Delta_{sr}, \varepsilon_{sr})$ are independent given Γ_s.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

This, combined with $X_{sri} = f_i(\Gamma_s + \Delta_{sr}, \varepsilon_{sr})$, implies LI-1 and LI-2 in the following way. For each (s, r), given θ_{sr}, the X_{sri} are a function of ε_{sri}. Because for each (s, r), the ε_{sri} are independent given θ_{sr}, the X_{sri} are independent given θ_{sr}, and LI-1 is implied. Furthermore, the $\Delta_{sr}, \varepsilon_{sri}$ are independent given Γ_s. Because X_{sri} are a function of $(\Gamma_s + \Delta_{sr}, \varepsilon_{sri})$, given Γ_s the X_{sri} are a function of $(\Delta_{sr}, \varepsilon_{sri})$. Hence, $\mathbf{X}_{sr}$ are independent given Γ_s, and LI-2 is implied. □

Proof of Theorem 2. *(Under B and H, UN-1, MO-1, and IIO-1 imply UN-2, MO-2, and IIO-2, respectively.)*

Proof. First, we consider the extreme case of no respondent variance: If $\text{var}(\Delta_{sr}) = 0$, then $\Delta_{sr} = 0$ and $\Theta_{sr} = \Gamma_s$ for all r and all s. As a result, $P(X_{sri} = x|\Theta_{sr}) = P(X_{sri} = x|\Gamma_s)$, $P(X_{sri} \geq x|\Theta_{sr}) = P(X_{sri} \geq x|\Gamma_s)$, and $E(X_{sri}|\Theta_{sr} = \theta_{sr}) = E(X_{sri}|\Gamma_s = \gamma_s)$. Hence, UN-1 = UN-2 (Definiton 1), MO-1 = MO-2 (Definition 3), and IIO-1 = IIO-2 (Definition 4). Second, for $\text{var}(\Delta_{sr}) > 0$, Lemma A1 proves that UN-1 implies UN-2, Lemma A2 proves that MO-1 implies MO-2, and Lemma A3 proves that IIO-1 implies IIO-2. □

Proof of Theorem 3. *(For dichotomous item scores (a) the MHM-1 implies MLR-1 and (b) for $R_s \geq I$, the MHM-2 implies MLR-2.)*

Proof.

(a) The assumptions in MHM-1 are identical to the assumptions used by Grayson (1988, Theorem 2) and Huynh (1994) to establish MLR of the sum score in Θ_{sr}, hence their proof can be applied.

(b) For clarity, we give the proof for $R = I$, but it can straightforwardly be generalized for $R > I$. Let $D = R!$ be the number of ways that respondents $1, \ldots, R$ can be ordered, and let d $(d = 1, \ldots, D)$ be an index of possible respondent orderings. Furthermore, let d_r $(r = 1, \ldots, R)$ denote the the position of respondent r in respondent-ordering d. For $R > I$, the same method can be applied, but each permutation contains only I respondents, hence the number of permutations $D = \frac{R!}{(R-I)!}$.

Let $X^d_{s+} = X_{sd_11} + X_{sd_22} + ... + X_{sd_rr} + ... + X_{sd_RR}$ denote the group-level sum score in which each item score is taken from a different respondent, with realization x^d_{s+}. Let $\mathbf{X}^d_s = (X^d_{s1}, X^d_{s2}, \ldots, X^d_{sI})$ denote the vector of item scores from the respondent order from the dth permutation, with realization $\mathbf{x}^d_s$.

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

For a given permutation *, let $\sum_{\{\mathbf{x}_s^* | \mathbf{x}_s^{*\prime}\mathbf{1}=x_{s+}^*\}}$ denote the sum over all possible patterns of I item scores that sum to x_{s+}^*. Let $P_i(\gamma_s) = P(X_{sri} = 1|\Gamma_s = \gamma_s)$ and let $Q_i(\gamma_s) = 1 - P_i(\gamma_s)$. By LI-2, for $r \neq p$, item scores X_{sri} and X_{spj} are independent conditional on γ_s. Hence, for dichotomous items, the probability of obtaining group-level sum score x_{s+}^* is

$$\begin{aligned} P(X_{s+}^* = x_{s+}^* | \Gamma_s = \gamma_s) &= \textstyle\sum_{\{\mathbf{x}_s^* | \mathbf{x}_s^{*\prime}\mathbf{1}=x_{s+}^*\}} \prod_{i=1}^{I} P_i(\gamma_s)^{x_{si}^*} Q_i(\gamma_s)^{(1-x_{si}^*)} \\ &= \textstyle\sum_{\{\mathbf{x}_s^* | \mathbf{x}_s^{*\prime}\mathbf{1}=x_{s+}^*\}} \prod_{i=1}^{I} Q_i(\gamma_s) [\frac{P_i(\gamma_s)}{Q_i(\gamma_s)}]^{x_{si}^*}. \end{aligned} \tag{A5}$$

Because $\prod_{i=1}^{I} Q_i(\gamma_s)$ is constant across each item-score pattern $\mathbf{x}$, Equation A5 is identical to

$$P(X_{s+}^* = x_{s+}^* | \Gamma_s = \gamma_s) = \textstyle\prod_{i=1}^{I} Q_i(\gamma_s) \sum_{\{\mathbf{x}_s^* | \mathbf{x}_s^{*\prime}\mathbf{1}=x_{s+}^*\}} \prod_{i=1}^{I} \left[\frac{P_i(\gamma_s)}{Q_i(\gamma_s)}\right]^{x_{si}^*}. \tag{A6}$$

The form of the right-hand side in Equation A6 is equal to the form used by Grayson (1988, Theorem 2) and Huynh (1994). Hence, their methods can be applied to establish MLR of the sum score X_{s+}^* in γ_s. Because $E(X_{si}^*|\Gamma_s = \gamma_s) = E(X_{sri}|\Gamma_s = \gamma_s)$,

$$\begin{aligned} E(X_{s+}^*|\Gamma_s = \gamma_s) &= \textstyle\sum_{i=1}^{I} E(X_{si}^*|\Gamma_s = \gamma_s) \\ &= \textstyle\sum_{i=1}^{I} E(X_{sri}|\Gamma_s = \gamma_s) \\ &= E(X_{s+}|\Gamma_s = \gamma_s), \end{aligned} \tag{A7}$$

it follows that MLR also holds for X_{s+} in Γ_s.

□

Proof of Theorem 4. *((a) The MHM-1 implies SOM-1 and (b) the MHM-2 implies SOM-2.)*

Proof.

(a) SOM-1 follows from the general result presented in Lemma A4. First, note that the respondent-level sum score X_{sr+} is a nondecreasing function of $\mathbf{X}_{sr}$ (e.g., Rosenbaum, 1984). Hence, by Lemma A4, X_{sr+} is nondecreasing in θ_{sr}, which is the definition of SOM-1.

(b) SOM-2 follows from the general result presented in Lemma A5. First, note that the group-level sum score X_{s+} is a nondecreasing function of $\mathbf{X}_s$. Hence, by Lemma A5, X_{s+} is nondecreasing in γ_s, which is the definition of SOM-2.

□

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

Proof of Theorem 5. *((a) The MHM-1 implies WSOL-1 and (b) the MHM-2 implies WSOL-2.)*

Proof.

(a) MHM-1 implies SOM-1 of X_{sr+} by Θ_{sr} (Theorem 4(a)). Van der Ark and Bergsma (2010, Theorem) showed that SOM implies WSOL, hence WSOL of Θ_{sr} by X_{sr+} is implied.

(b) Similar to the proof in (a), the MHM-2 implies SOM-2 of X_{s+} by Γ_s, hence, WSOL of Γ_s by X_{s+} is implied.

□

Proof of Theorem 6. *(For dichotomous items (a) the MHM-1 implies MM-1 and (b) the MHM-2 implies MM-2.)*

Proof.

(a) The proof is analogous to the proof in Proposition 4.1a Junker (1993). By the law of total expectation (e.g., Rice, 2006, p 149) and LI-1

$$\begin{aligned} E(X_{sri}|X_{sr(i)}) &= E[E(X_{sri}|X_{sr(i)},\Theta_{sr})|X_{sr(i)}] \\ &= E[E_i(\Theta_{sr})|X_{sr(i)}]. \end{aligned} \tag{A8}$$

For dichotomous items, under the MHM-1, by Theorem 3(a), Θ_{sr} is nondecreasing in $X_{sr(i)}$ (SOL). Because Θ_{sr} is stochastically ordered in $X_{sr(i)}$, so is any nondecreasing function of Θ_{sr}, such as $E_i(\Theta_{sr})$ (Equation 6; Shaked & Shanthikumar, 2007, Theorem 1.A.3.(a)), which completes the proof.

(b) This proof is parallel to the proof in (a), which holds when substituting Θ_{sr} by Γ_s, $X_{sr(i)}$ by $X_{s(r,i)}$, $E_i(\cdot)$ by $\mathcal{E}(\cdot)$, LI-1 by LI-2, and MHM-1 by MHM-2.

□

Proof of Theorem 7. *((a) The MHM-1 implies CA-1 and (b) the MHM-2 implies CA-2.)*

Proof.

(a) The proof is similar to the proof of Theorem 1 by (Rosenbaum, 1984; see also Holland & Rosenbaum, 1986, Theorem 6). If CA-1 holds, the conditional covariance $Cov[g_1(\mathbf{Y}_{sr}), g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] \geq 0$ (Definition 9). Using standard algebra, it can be shown that this statement is equivalent to

$$E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] \geq E[g_1(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})]E[g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] \tag{A9}$$

(e.g., Rice, 2006, p. 138). Hence, we prove that under the MHM-1 Equation A9 holds. By the law of total expectation,

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

$$E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] = E\{E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}), \theta]|h(\mathbf{Z}_{sr})\} \quad \text{(A10)}$$

(Rice, 2006, p. 138). By LI-1, $\mathbf{Y}_{sr}$ and $\mathbf{Z}_{sr}$ are independent given θ_{sr}. Hence,

$$E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] = E\{E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|\theta_{sr}]|h(\mathbf{Z}_{sr})\}. \quad \text{(A11)}$$

Because, by LI-1, the values in $\mathbf{Y}_{sr}$ are independent given θ_{sr}, they are associated (Esary et al., 1967, Theorem 2.1). Therefore,

$$E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] \geq E\{E[g_1(\mathbf{Y}_{sr})|\theta_{sr}]E[g_2(\mathbf{Y}_{sr})|\theta_{sr}]|h(\mathbf{Z}_{sr})\}. \quad \text{(A12)}$$

By Lemma A4, $E(g_1(\mathbf{Y}_{sr}|\theta_{sr})$ and $E(g_2(\mathbf{Y}_{sr}|\theta_{sr})$ are nondecreasing in θ_{sr}, hence, they are associated (Esary et al., 1967, P_4). In addition, by UN-1, θ_{sr} is a scalar and therefore associated (Esary et al., 1967, P_3). Hence, it follows that

$$E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] \geq E\{E[g_1(\mathbf{Y}_{sr})|\theta_{sr}]|h(\mathbf{Z}_{sr})\}E\{E[g_2(\mathbf{Y}_{sr})|\theta_{sr}]|h(\mathbf{Z}_{sr})\}. \quad \text{(A13)}$$

By the law of total expectation, the statement in Equation A13 is equivalent to

$$E[g_1(\mathbf{Y}_{sr})g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})] \geq E[g_1(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})]E[g_2(\mathbf{Y}_{sr})|h(\mathbf{Z}_{sr})], \quad \text{(A14)}$$

which completes the proof.

(b) The proof is similar to the proof in (a). Throughout the proof, $g_1(\mathbf{Y}_{sr})$ is kept the same, but $g_2(\mathbf{Y}_{sr})$ is replaced by $g_2(\mathbf{Y}_{sp})$, with $r \neq p$. Hence, g_1 and g_2 apply to different respondents within the same group. Furthermore, $\mathbf{Z}_{sr}$ is replaced by $\mathbf{Z}_{s(rp)}$, hence to the vector that contains all item scores in group s, except the scores of respondents r and p. Finally, θ_{sr} is replaced by γ_s, LI-1 by LI-2, and Lemma A4 by Lemma A5, which gives the proof for (b) (see, also, Rosenbaum, 1988).

□

Proof of Theorem 8. *((a) The DMM-1 implies MIIO-1 and (b) the DMM-2 implies MIIO-2.)*

Proof.

(a) This proof is similar to the proof of the Corollary by Ligtvoet et al. (2011). By IIO-1, $E(X_{sri}|\Theta_{sr} = \theta_{sr}) \leq E(X_{srj}|\Theta_{sr} = \theta_{sr})$. By Equation 5, $E(X_{sri}|\Theta_{sr} = \theta_{sr}) \leq E(X_{srj}|\Theta_{sr} = \theta_{sr})$ equals

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press

$$\begin{aligned} & \textstyle\sum_{x=1}^{m} P(X_{sri} \geq x|\Theta_{sr} = \theta_{sr}) \;\; \leq \;\; \sum_{x=1}^{m} P(X_{srj} \geq x|\Theta_{sr} = \theta_{sr}) \\ \Leftrightarrow \;\; & \textstyle\sum_{x=1}^{m} P(X_{sri} \geq x|\Theta_{sr} = \theta_{sr}) P(X_{sr(ij)} = y|\Theta_{sr} = \theta_{sr}) \leq \\ & \qquad \textstyle\sum_{x=1}^{m} P(X_{srj} \geq x|\Theta_{sr} = \theta_{sr}) P(X_{sr(ij)} = y|\Theta_{sr} = \theta_{sr}). \end{aligned} \tag{A15}$$

By LI-1, X_{sri} and $X_{sr(ij)}$ are independent given θ_{sr}, and their joint probability equals the product of their marginal conditional probabilities (e.g., Rice, 2006, p. 84). Hence, Equation A15 equals

$$\textstyle\sum_{x=1}^{m} P(X_{sri} \geq x, X_{sr(ij)} = y|\Theta_{sr} = \theta_{sr}) \;\; \leq \;\; \sum_{x=1}^{m} P(X_{srj} \geq x, X_{sr(ij)} = y|\Theta_{sr} = \theta_{sr}). \tag{A16}$$

Let $F(\Theta_{sr})$ denote the cumulative distribution function of Θ_{sr}. Integrating both sides of Equation A16 over Θ_{sr} yields

$$\begin{aligned} & \textstyle\int \sum_{x=0}^{m} P(X_{sri} \geq x, X_{sr(ij)} = y|\Theta_{sr} = \theta_{sr}) dF(\Theta_{sr}) \;\; \leq \\ & \qquad\qquad \textstyle\int \sum_{x=0}^{m} P(X_{sri} \geq x, X_{sr(ij)} = y|\Theta_{sr} = \theta_{sr}) dF(\Theta_{sr}) \\ \Leftrightarrow \;\; & \textstyle\sum_{x=1}^{m} P(X_{sri} \geq x, X_{sr(ij)} = y) \;\; \leq \;\; \sum_{x=1}^{m} P(X_{srj} \geq x, X_{sr(ij)} = y) \\ \Leftrightarrow \;\; & \textstyle\sum_{x=1}^{m} P(X_{sri} \geq x|X_{sr(ij)} = y) \;\; \leq \;\; \sum_{x=1}^{m} P(X_{srj} \geq x|X_{sr(ij)} = y) \\ \Leftrightarrow \;\; & E(X_{sri}|X_{sr(ij)}) \;\; \leq \;\; E(X_{srj}|X_{sr(ij)}), \end{aligned} \tag{A17}$$

for all y and all $i < j$, completing the proof.

(b) The proof is parallel to the proof in (a). Replacing Θ_{sr} by Γ_s, θ_{sr} by γ_s, IIO-1 by IIO-2, LI-1 by LI-2, $X_{sr(ij)}$ by $X_{s(r,ij)}$, proofs that the MHM-2 implies MIIO-2. □

https://doi.org/10.1017/psy.2024.9 Published online by Cambridge University Press