PSYCHOMETRIKA—VOL. 86, NO. 2, 378–403
JUNE 2021
https://doi.org/10.1007/s11336-021-09762-5

MAPPING UNOBSERVED ITEM–RESPONDENT INTERACTIONS: A LATENT SPACE ITEM RESPONSE MODEL WITH INTERACTION MAP

MINJEONG JEON

UNIVERSITY OF CALIFORNIA, LOS ANGELES

ICK HOON JIN

YONSEI UNIVERSITY

MICHAEL SCHWEINBERGER

RICE UNIVERSITY

SAMUEL BAUGH

UNIVERSITY OF CALIFORNIA, LOS ANGELES

Classic item response models assume that all items with the same difficulty have the same response probability among all respondents with the same ability. These assumptions, however, may very well be violated in practice, and it is not straightforward to assess whether these assumptions are violated, because neither the abilities of respondents nor the difficulties of items are observed. An example is an educational assessment where unobserved heterogeneity is present, arising from unobserved variables such as cultural background and upbringing of students, the quality of mentorship and other forms of emotional and professional support received by students, and other unobserved variables that may affect response probabilities. To address such violations of assumptions, we introduce a novel latent space model which assumes that both items and respondents are embedded in an unobserved metric space, with the probability of a correct response decreasing as a function of the distance between the respondent's and the item's position in the latent space. The resulting latent space approach provides an interaction map that represents interactions of respondents and items, and helps derive insightful diagnostic information on items as well as respondents. In practice, such interaction maps enable teachers to detect students from underrepresented groups who need more support than other students. We provide empirical evidence to demonstrate the usefulness of the proposed latent space approach, along with simulation results.

Key words: item response data, latent space model, network model, bipartite network, interactions, interaction map.

1. Introduction

Item response theory (IRT) is a widely used approach for analyzing responses to test items given by test takers, called respondents. A classic IRT model, the Rasch model (Rasch 1961), assumes that the log odds of the probability of a correct response $Y_{j,i} = 1$ to binary item i by respondent j is of the form:

$$\text{logit}(\mathbb{P}(Y_{j,i} = 1 \mid \alpha_j,\ \beta_i)) \ = \ \alpha_j + \beta_i. \tag{1}$$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-021-09762-5.

Minjeong Jeon, Ick Hoon Jin, and Michael Schweinberger are co-first authors with equal contribution.

Correspondence should be made to Minjeong Jeon, UNIVERSITY OF CALIFORNIA, LOS ANGELES, Los Angeles, CA90095, USA. Email: mjjeon@ucla.edu

In words, the probability of a correct response to item i by respondent j is a function of two attributes: one associated with respondent j, $\alpha_j \in \mathbb{R}$, and the other associated with item i, $\beta_i \in \mathbb{R}$. The main effect α_j represents the ability of respondent j, while the main effect β_i of item i reveals how easily item i is correctly answered.

The Rasch model rests on the following assumptions: (1) The item responses of any respondent are independent of the item responses of any other respondent, conditional on the abilities of the respondents and the easiness levels of the items; (2) for each respondent, the responses to items are independent, conditional on the ability of the respondent and the easiness levels of the items; (3) for each item, respondents with the same level of ability have the same success probability; and for each respondent, items with the same easiness level have the same success probability.

These assumptions, however, may very well be violated in practice: For example, in some educational assessments, it is not credible that all items with the same easiness level have the same response probability for all respondents with the same ability. An example is an educational assessment where unobserved heterogeneity is present, arising from unobserved variables such as cultural background and upbringing of students, the quality of mentorship and other forms of emotional and professional support received by students, and other unobserved variables that may affect response probabilities. Worse, in practice it is not straightforward to assess whether such assumptions are violated, because the abilities of respondents and the easiness levels of items are unobserved.

To address violations of those assumptions, we introduce a novel latent space model which assumes that both items and respondents are embedded in an unobserved metric space, with the probability of a correct response decreasing as a function of the distance between the respondent's and the item's position in the latent space. The resulting latent space approach provides an interaction map that represents interactions of respondents and items, and helps derive insightful diagnostic information on items as well as respondents.

The novel latent space model we introduced is inspired by recent work on item response models that view item response data as networks. For example, Borsboom (2008) described a network analysis of psychological constructs, where covariance between observed indicator variables stems from interactions among items. More recently, Epskamp et al.(2018) proposed a network approach based on Gaussian graphical models, which can include latent variables. Marsman et al. (2018) studied relations between an Ising model and other item response models. All of them are concerned with interactions among items, not respondents. Another recent development in network modeling of item response data is the doubly latent space joint model (DLSJM) of Jin and Jeon (2019) and its extension to hierarchical data (Jin et al. 2018), henceforth called the network item response model (NIRM). The NIRM approach is inspired by latent space models of network data (Hoff et al. 2002; Schweinberger and Snijders 2003; Sewell and Chen 2015; Smith et al. 2019). Latent space models of network data and non-network data may be viewed as a model-based alternative to multidimensional scaling (MDS), having the advantage of enabling model-based statistical inference and capturing the uncertainty about the positions of units in the latent space, in contrast to MDS. The NIRM approach constructs functions of item response data which can be viewed as network data: respondent-respondent networks consisting of links between respondents who both gave the correct response to an item (one network for each item), and item-item networks consisting of links between items that received the correct response by a respondent (one network for each respondent). Our proposed approach is inspired by NIRM, but simpler than the NIRM approach. We view item response data as a bipartite network, consisting of links between respondents, on the one hand, and items, on the other hand. This change in perspective comes with important benefits, including, but not limited to: (1) We work with the original item response data rather than functions of item response data; (2) we have a single network rather than multiple networks, which would have to be combined; (3) we can examine relationships between items and respondents without choosing a procedure that combines multiple networks, which—when the procedure is inappropriate—introduces an additional source of error; and (4) our approach is closely related to the Rasch model, which facilitates interpretation.

Our paper is organized as follows: We introduce latent space models in Sect. 2, discuss Bayesian inference in Sect. 3, and present examples along with simulation results in Sects. 4 and 5. We conclude our paper in Sect. 6.

2. Model

2.1. Latent Space Item Response Model

We consider item response data consisting of a binary N by I matrix $\boldsymbol{Y} \in \{0, 1\}^{N\times I}$, where $Y_{j,i} = 1$ indicates a correct response by respondent j to item i, whereas $Y_{j,i} = 0$ indicates an incorrect response. Extensions to non-binary item response data are straightforward, by replacing the logit-link function for binary item response data by a suitable link function for non-binary item response data, as in generalized linear models (McCullagh and Nelder 1983).

To capture unobserved interactions of respondents and items, we assume that both respondents and items are embedded in an unobserved metric space. A metric space $(\mathbb{M}, d)$ consists of a space $\mathbb{M}$ and a distance function $d : \mathbb{M} \times \mathbb{M} \mapsto [0, +\infty)$ assigning distances to pairs of points $(\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{M} \times \mathbb{M}$ (corresponding to positions of respondents and items), which satisfy

- reflexivity: $d(\boldsymbol{a},\ \boldsymbol{b}) = 0$ if and only if $\boldsymbol{a} = \boldsymbol{b} \in \mathbb{M}$;
- symmetry: $d(\boldsymbol{a},\ \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a})$ for all $\boldsymbol{a},\ \boldsymbol{b} \in \mathbb{M}$;
- triangle inequality: $d(\boldsymbol{a},\ \boldsymbol{b}) \leq d(\boldsymbol{a}, \boldsymbol{c}) + d(\boldsymbol{b}, \boldsymbol{c})$ for all $\boldsymbol{a},\ \boldsymbol{b},\ \boldsymbol{c} \in \mathbb{M}$.

We follow the convention in statistical network analysis (Hoff et al. 2002) and assume that $\mathbb{M}$ is p-dimensional Euclidean space $\mathbb{R}^p$ with known dimension $p \geq 1$. Some possible choices of the distance function $d : \mathbb{R}^p \times \mathbb{R}^p \mapsto [0, +\infty)$ are:

- ℓ_1-distance (city-block distance): $d(\boldsymbol{a},\ \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}||_1 = \sum_{i=1}^{p} |a_i - b_i|$,
- ℓ_2-distance (Euclidean distance): $d(\boldsymbol{a},\ \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}||_2 = \sqrt{\sum_{i=1}^{p}(a_i - b_i)^2}$,
- ℓ_∞-distance (maximum distance): $d(\boldsymbol{a},\ \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}||_\infty = \max_{1\leq i\leq p}\ |a_i - b_i|$.

To capture unobserved interactions of respondents and items, we assume that the probability of a correct response by respondent j to item i depends on the position $\boldsymbol{a}_j \in \mathbb{R}^p$ of respondent j and the position $\boldsymbol{b}_i \in \mathbb{R}^p$ of item i in the shared metric space:

$$\operatorname{logit}(\mathbb{P}(Y_{j,i} = 1 \mid \alpha_j,\ \beta_i,\ \boldsymbol{a}_j,\ \boldsymbol{b}_i)) \ = \ \alpha_j + \beta_i + g(\boldsymbol{a}_j,\ \boldsymbol{b}_i), \tag{2}$$

where $g : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ is a real-valued function of the positions of respondent j and item i. There are many possible choices of the function g. We discuss two natural choices:

- multiplicative effect: $g(\boldsymbol{a}_j,\ \boldsymbol{b}_i) = \boldsymbol{a}_j^\top \boldsymbol{b}_i$, where $\boldsymbol{a}_j^\top \boldsymbol{b}_i$ is the inner product of $\boldsymbol{a}_j$ and $\boldsymbol{b}_i$;
- distance effect: $g(\boldsymbol{a}_j,\ \boldsymbol{b}_i) = -\gamma\ d(\boldsymbol{a}_j, \boldsymbol{b}_i)$, where $d(\boldsymbol{a}_j,\ \boldsymbol{b}_i)$ is the distance between $\boldsymbol{a}_j$ and $\boldsymbol{b}_i$ (e.g., the ℓ_1-distance, ℓ_2-distance, or ℓ_∞-distance) and $\gamma \geq 0$ is the weight of the distance term; note that $\gamma > 0$ ensures that increasing the distance decreases the probability of a correct response.

While both choices are legitimate and have advantages and disadvantages, we believe that the distance effect is easier to interpret than the multiplicative effect. For example, the effect of the inner product on the log odds of a correct response is 0 when $\boldsymbol{a}_j$ and $\boldsymbol{b}_i$ are orthogonal, regardless of whether the distance between $\boldsymbol{a}_j$ and $\boldsymbol{b}_i$ is small or large: For example, if d is the ℓ_2-distance and $\boldsymbol{a}_j = (0,\ 1/100)$ and $\boldsymbol{b}_i = (1/100,\ 0)$, then $d(\boldsymbol{a}_j,\ \boldsymbol{b}_i) = 0.01$, whereas $\boldsymbol{a}_j = (0,\ 100)$ and $\boldsymbol{b}_i = (100,\ 0)$ imply $d(\boldsymbol{a}_j,\ \boldsymbol{b}_i) = 141.42$. In both examples, $\boldsymbol{a}_j^\top \boldsymbol{b}_i = 0$, but in the first case the distance between the two vectors is small, whereas in the second case, it is large. Therefore, to interpret interaction maps and the effect of interactions on the probability of a correct response under the multiplicative effects model, one needs to pay careful attention to the angle of the vectors $\boldsymbol{a}_j$ and $\boldsymbol{b}_i$, in addition to the lengths of $\boldsymbol{a}_j$ and $\boldsymbol{b}_i$. That makes the resulting interaction maps more challenging to use by practitioners and applied researchers, undermining one of the main advantages of the latent space approach. We therefore focus on the model with the distance effect, although the multiplicative effects model would be an interesting alternative.

It is worth noting that the latent space model with $\gamma = 0$ is equivalent to the Rasch model, so the latent space model with $\gamma \geq 0$ can be viewed as a generalization of the Rasch model. In practice, we determine whether $\gamma = 0$ or $\gamma > 0$ via model selection, as described in Sect. 3.3. If $\gamma > 0$, the latent space model has added value compared with the Rasch model. The added value of the latent space model is that it captures deviations from the main effects α_j and β_i of the Rasch model—that is, interactions of respondent j and item i—and visualizes those interactions by embedding respondents along with items in a shared metric space. As a consequence, it is natural to interpret the metric space as an interaction map, rather than an ability space.

We discuss below properties of the latent space model, including practical and theoretical advantages along with a network view of item response data. Statistical issues—including identifiability issues—are discussed in Sect. 3.

2.2. *Properties*

2.2.1. Latent Space Model as Network Model The latent space model introduced above was inspired by latent space models of network data—as mentioned in Sect. 1—and it may be viewed as a network model. Specifically, one may view item response data as a bipartite network (Wasserman and Faust 1994), consisting of links between respondents, on the one hand, and items, on the other hand, where links correspond to correct responses by respondents to items. In contrast to conventional network data, bipartite respondent–item networks consist of two sets of units rather than one set of units (i.e., the set of respondents and the set of items). The proposed modeling framework then assumes that the bipartite network was generated by a latent space model, with both respondents and items embedded in a shared metric space.

It is worth noting that viewing the proposed latent space model as a network model may or may not be useful, although network models have turned out to be useful across a staggering number of fields—not the least in artificial intelligence (AI), where deep neural networks have enabled substantial advances in voice recognition and computer vision (Goodfellowet al. 2016). In probability and statistics, there are two main lines of research involving networks. First, graphical models use graphs to represent conditional independence structure (Pearl 1988; Lauritzen 1996), that is, model structure. Graphical models assume that the variables of interest (here: items) are the vertices of a graph, and the absence of a link between two vertices indicates that the two corresponding variables are independent conditional on all other variables. Links therefore indicate conditional dependencies, given all other vertices. Examples are Gaussian graphical models, Ising models, and Boltzmann machines in AI. In the IRT literature, Epskamp et al. (2018) and Marsman et al. (2018) and others followed a graphical model approach to studying interactions among items (albeit not respondent–item interactions, as we do). Second, random graph models use graphs to represent data structure. For example, in social network analysis (Wasserman and Faust 1994), individuals are the vertices of a graph, and the links may indicate friendships among individuals.

The proposed latent space model can be represented as a graphical model (with variables $Y_{j,i}$, $\boldsymbol{a}_j$, and $\boldsymbol{b}_i$ constituting the vertices of a graph and links corresponding to conditional dependencies among these variables) or as a random graph model (with respondents j and items i constituting

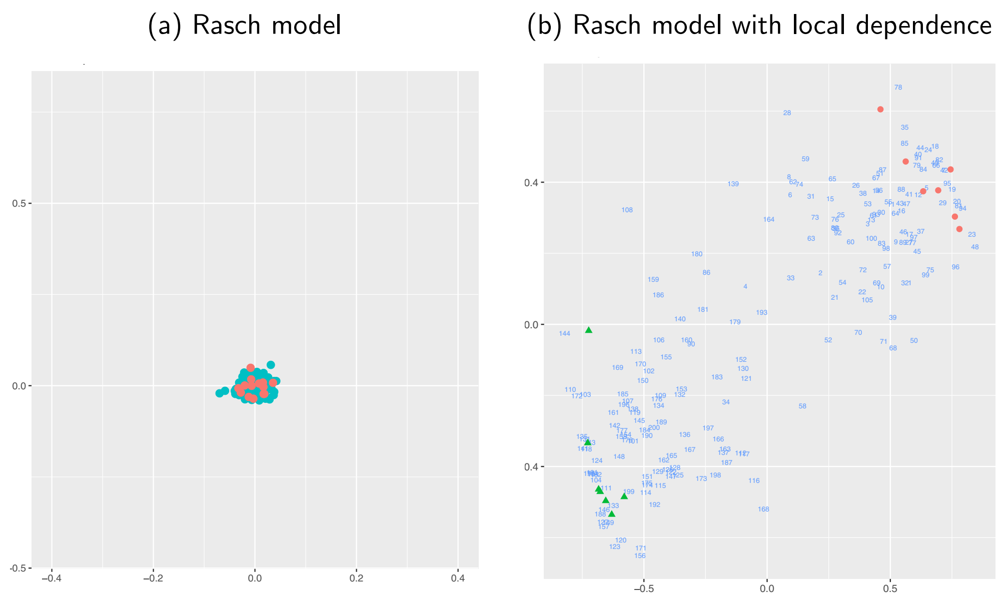

FIGURE 1.
Estimated latent space configurations based on responses to 14 items by 200 respondents, generated by **a** the Rasch model and **b** the Rasch model with local dependence. In **a**, blue circles represent respondents and red circles represent items, whereas in **b** blue numbers represent respondents, red circles represent the first 7 items, and green triangles represent the last 7 items (Color figure online)

the vertices of a graph and links corresponding to correct responses). Whether it is useful to view the proposed model as a graphical model or as a random graph model is open to discussion, but—regardless of whether one embraces a network view—the proposed modeling framework has practical and theoretical advantages.

2.2.2. Practical Advantages A unique advantage of the proposed latent space approach is that it provides a geometric representation of interactions among respondents and items in a low-dimensional space, e.g., $\mathbb{R}^2$. The interaction structure mapped into two-dimensional Euclidean space helps detect unobserved characteristics of items and respondents.

To demonstrate, we conduct a simulation study. The simulation results are based on binary responses to 14 items by 200 respondents. First, data are generated from the Rasch model. Second, data are generated from the Rasch model with local dependence, that is, the responses of the first 100 respondents to the first 7 items exhibit strong local dependence in the sense of Chen and Thissen (1997), and the responses of the last 100 respondents to the last 7 items likewise exhibit strong local dependence. The proposed latent space model with two-dimensional Euclidean space $\mathbb{R}^2$ is estimated from both datasets, using the Bayesian Markov chain Monte Carlo algorithm described in Sect. 3. Additional details are provided in "Appendix A" of the supplement. Figure 1 shows that in the first case all items and respondents are located close to the origin of $\mathbb{R}^2$, whereas in the second case the two groups of items are well separated in $\mathbb{R}^2$ and the two groups of 100 respondents are located close to the respective sets of items, as expected.

Latent space dimension Throughout the remainder of the paper, we choose $\mathbb{M} = \mathbb{R}^2$, because a two-dimensional space has clear advantages in terms of parsimony, ease of interpretability, and visualization. As mentioned above, it is natural to interpret $\mathbb{R}^2$ as an interaction map rather than an ability map, because the added value of the latent space model is that it captures deviations

TABLE 1.
Hypothetical item response matrix consisting of four respondents 1, 2, 3, 4 and six items I1, I2, I3, I4, I5, I6

Response	I1	I2	I3	I4	I5	I6
1	1	1	1	0	0	0
2	1	1	1	0	0	0
3	0	0	0	1	1	1
4	0	0	0	1	1	1

The four respondents show two response patterns. Respondents 1 and 2 give correct responses to Items I1–I3 only, whereas Respondents 3 and 4 give correct responses to Items I4–I6 only

from the main effects α_j and β_i of the Rasch model—that is, interactions of respondent j and item i—and visualizes those interactions by embedding respondents along with items in $\mathbb{R}^2$.

2.2.3. Theoretical Advantages Among the theoretical advantages of the proposed latent space model is the fact that it weakens the conditional independence assumptions of conventional IRT models, along with the homogeneity assumptions of classic IRT models.

Conditional independence assumptions

The proposed latent space model is based on the following conditional independence assumption:

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \gamma,\ \boldsymbol{A},\ \boldsymbol{B}) = \prod_{j=1}^{N} \prod_{i=1}^{I} \mathbb{P}(Y_{j,i} = y_{j,i} \mid \alpha_j,\ \beta_i,\ \gamma,\ \boldsymbol{a}_j,\ \boldsymbol{b}_i),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)$, $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)$, and $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_I)$. In words, the item responses are assumed to be independent conditional on the positions of respondents and items in the latent space, and the respondent and item attributes. This conditional independence assumption is weaker than the conditional independence of the Rasch model, which requires that item responses are independent conditional on respondent and item attributes. So the latent space model relaxes the conditional independence assumptions of the Rasch model and other classic IRT models.

The weaker conditional independence assumption of the latent space model allows for respondent–item interactions. As a consequence, the latent space model can account for local dependence among item responses arising from a variety of sources, including testlets (e.g., items similar in content), learning and practice effects, or repeated measurements, as well as person dependence stemming from shared school or family memberships (or even unobserved memberships).

Homogeneity assumptions

In addition, the latent space model drops some of the homogeneity assumptions made by conventional IRT models. For example, consider two respondents j_1 and j_2 with identical abilities, who are located at distances $d(\boldsymbol{a}_{j_1},\ \boldsymbol{b}_i) < d(\boldsymbol{a}_{j_2},\ \boldsymbol{b}_i)$ from item i. Then, respondent j_1 has a higher probability of giving a correct response to item i than respondent j_2, despite the fact that j_1 and j_2 have identical abilities. A similar scenario arises when two items i_1 and i_2 have identical difficulty levels and distances $d(\boldsymbol{a}_j,\ \boldsymbol{b}_{i_1}) > d(\boldsymbol{a}_j,\ \boldsymbol{b}_{i_2})$ from some respondent j, which implies that item i_1 is less likely to be answered correctly than item i_2, despite identical difficulty levels.

To give a specific example of an educational assessment where such homogeneity assumptions are violated and to demonstrate how the latent space approach captures those violations, we

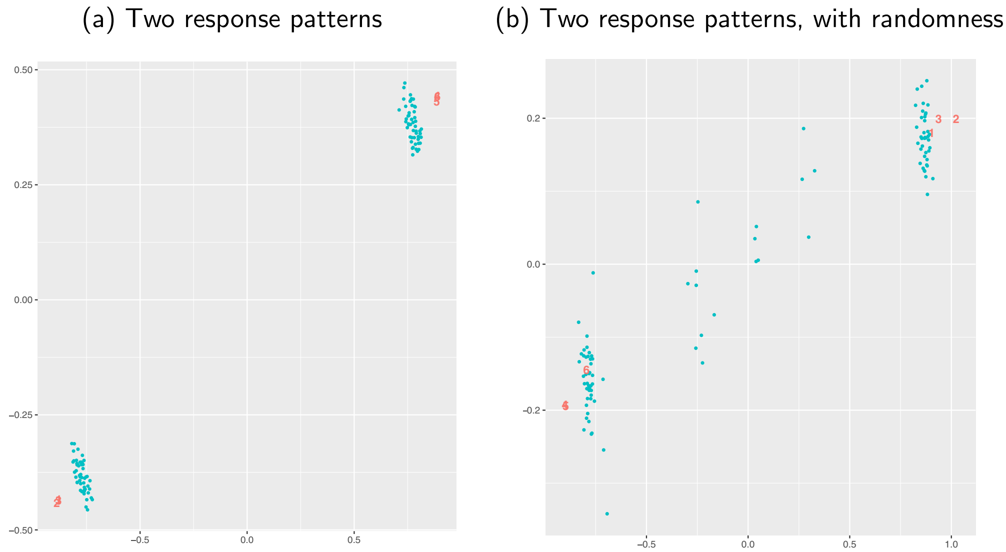

FIGURE 2.
a Latent space estimated from a hypothetical data example where Respondents 1–50 (bottom left) give correct responses to Items 1–3 only, whereas Respondents 51–100 (top right) give correct responses to Items 4–6 only. γ was estimated as 4.45 (posterior median, with 95% posterior credible interval [3.79, 5.20]). **b** Latent space estimated from a hypothetical setting similar to **a**: Respondents 1–40 (bottom left) give correct responses to Items 1–3, while Respondents 41–80 (top right) give correct responses to Items 4–6. The remaining 20 respondents (middle) give random responses to Items 1–6. γ was estimated as 3.62 [3.07, 4.26]. In both figures, blue circles represent respondents and red numbers represent items (Color figure online)

provide a hypothetical item response matrix in Table 1 with six math test items answered by four respondents. Suppose the first three items are algebra items, while the last three items are geometry items. Assume that the algebra and geometry items have identical difficulty levels. Table 1 shows that Persons 1 and 2 have all algebra item correct but none of the geometry items, whereas Persons 3 and 4 have all geometry items correct but no algebra item. In other words, all respondents have three correct responses, which is an indication that all respondents have similar abilities because the difficulties of the items are the same (by assumption). However, despite similar abilities of the four respondents and identical difficulty levels of the items, it is hard to believe that the response probabilities of all respondents and all items are similar.

Figure 2a represents estimated latent space configurations based on item response data mimicking the item response matrix in Table 1, with 50 respondents giving correct responses to the first three items but incorrect responses to the last three items, while the other 50 respondents give incorrect responses to the first three items and correct responses to the last three items. Figure 2b represents estimated latent space configurations based on item response data including random response patterns—80 respondents have response patterns similar to the response matrix in Table 1, while the other 20 respondents give random responses to Items 1–6. Figure 2 reveals that the latent space approach separates the two groups of respondents in both cases, with and without randomness. While the scenario in Fig. 2a may be an extreme-case scenario, the scenario in Fig. 2b is more realistic, and may very well be encountered in practice.

2.3. *Related Models*

We review related models, excluding those we have already reviewed in Sect. 1.

2.3.1. Other Models with Relaxed Assumptions As discussed above, the proposed latent space model weakens the assumptions of classic IRT models, allowing for unobserved heterogeneity and dependence among item responses. Other approaches to relaxing those assumptions include polytomous item models, testlet and bifactor models, interaction effects models (e.g., Wainer and Kiely 1987; Wilson and Adams 1995), finite mixture models with latent classes (e.g., Rost 1990,), and multilevel models (e.g., Fox and Glas 2001,).

Although these approaches have been applied successfully in applications, they are not free of limitations. For instance, many of them require the dependence structure of items and respondents to be known prior to data analysis, which is a strong assumption. The latent space model does not require knowledge of the interaction structure. In addition, the discussed approaches relax some but not all of the assumptions: For example, finite mixture models allow for heterogeneity between latent classes, but assume homogeneity within latent classes. In addition, finite mixture models assume that there are latent classes, which is equivalent to assuming that there is an unobserved, discrete metric space. By contrast, the proposed latent space model assumes that the unobserved metric space is continuous rather than discrete, offering more flexibility to represent respondent–item interactions.

2.3.2. Other Models with Interactions Among Respondents and Items We discussed that the proposed latent space model can be viewed as a generalization of the Rasch model with an additional distance term. In the sense that distances represent relations between respondents and items, after controlling for the main effects of respondent and item attributes, distances can be regarded as interactions between respondents and items that are not explained with the main effects.

Two-parameter IRT model

An alternative model that captures interactions among respondents and items is the two-parameter IRT model, which assumes that

$$\text{logit}(\mathbb{P}(Y_{j,i} = 1 \mid \alpha_j,\ \beta_i,\ \lambda_i)) \ = \ \lambda_i\ \alpha_j + \beta_i, \tag{3}$$

where $\lambda_i \in \mathbb{R}$. The term $\lambda_i\ \alpha_j$ captures interactions of item i and respondent j. The latent space approach has an important advantage over the two-parameter IRT model: It embeds both respondents and items in a low-dimensional space, helping visualize interactions of respondents and items.

Interaction IRT model

A more general interaction model assumes that

$$\text{logit}(\mathbb{P}(Y_{j,i} = 1 \mid \alpha_j,\ \beta_i,\ \epsilon_{j,i})) \ = \ \alpha_j + \beta_i + \epsilon_{j,i}, \tag{4}$$

where $\epsilon_{j,i} \in \mathbb{R}$ represents the interaction of respondent j and item i. The latent space model can be viewed as a special case of the interaction model, corresponding to

$$\epsilon_{j,i} \ = \ -\gamma\ d(\boldsymbol{a}_j,\ \boldsymbol{b}_i).$$

In other words, the latent space model makes the implicit assumption that the interaction effects $\epsilon_{j,i}$ are of the form $\epsilon_{j,i} = -\gamma\ d(\boldsymbol{a}_j,\ \boldsymbol{b}_i)$, where the distances $d(\boldsymbol{a}_j,\ \boldsymbol{b}_i)$ satisfy reflexivity, symmetry, and the triangle inequality, as described in Sect. 2.1. While the proposed latent space model is a special case of the interaction model, it has three advantages over the interaction model. First, the latent space model can be estimated, whereas the interaction model cannot be estimated unless additional parameter constraints are imposed, because in practice we have a single observation (i.e., item response) for each pair of respondents and items. Second, the latent space model captures transitivity in item response data thanks to the triangle inequality: For example, if the positions of two respondents j_1 and j_2 are close to the position of item i in the latent space, then the two respondents are fairly close to each other, by the triangle inequality; likewise, if two items i_1 and i_2 are close to respondent j, then the two items are fairly close to each other. The assumption that item response data are transitive makes sense in applications. Therefore, while the latent space model is more restrictive than the general interaction model, the restrictions make sense in practice, and facilitate estimation. Last, but not least, the latent space approach provides an interaction map of respondents and items.

Bilinear mixed effects models and related models

The multiplicative effects version of the latent space model is related to the bilinear mixed effects model of Hoff (2005), the additive and multiplicative effects models of Hoff (2020), and the latent factor models of Agarwal and Chen (2009). For example, the bilinear mixed effects models of Hoff (2005) are models of network data, such as friendships among N students. Bilinear mixed effects models add a multiplicative effect of the form $\boldsymbol{a}_j^\top \boldsymbol{b}_i$ to the log odds of a friendship between students i and j. The multiplicative effects version of the proposed latent space models resembles the multiplicative effects in the above-mentioned models, but multiplicative effects are more difficult to interpret, as pointed out in Sect. 2.

Differential item functioning

Last, but not least, IRT models for studying differential item functioning (DIF) can be seen as special cases of interaction models, where an interaction term is formed with a known categorical attribute of respondents (e.g., gender) and an item indicator. Conventional DIF models, however, require pre-knowledge of the respondent attribute. The proposed latent space model does not require such pre-knowledge.

3. Bayesian Inference

3.1. Markov Chain Monte Carlo (MCMC)

We propose a fully Bayesian approach for estimating the proposed latent space model, using MCMC methods. Bayesian inference is preferable to maximum likelihood due to under-identification of the latent space positions.

We use the following priors:

$$
\begin{array}{lll}
\alpha_j \mid \sigma^2 & \overset{\text{ind}}{\sim} \mathrm{N}\left(0,\ \sigma^2\right), & \sigma^2 > 0, \quad j = 1, \ldots, N \\
\beta_i \mid \tau_\beta^2 & \overset{\text{ind}}{\sim} \mathrm{N}\left(0,\ \tau_\beta^2\right), & \tau_\beta^2 > 0, \quad i = 1, \ldots, I \\
\log \gamma \mid \mu_\gamma,\ \tau_\gamma^2 & \sim \mathrm{N}\left(\mu_\gamma,\ \tau_\gamma^2\right), & \mu_\gamma \in \mathbb{R}, \quad \tau_\gamma^2 > 0 \\
\sigma^2 \mid a_\sigma,\ b_\sigma & \sim \text{Inv-Gamma}\left(a_\sigma,\ b_\sigma\right), & a_\sigma > 0, \quad b_\sigma > 0 \\
\boldsymbol{a}_j & \overset{\text{iid}}{\sim} \mathrm{MVN}_p\left(\mathbf{0},\ \boldsymbol{I}_p\right), & j = 1, \ldots, N \\
\boldsymbol{b}_i & \overset{\text{iid}}{\sim} \mathrm{MVN}_p\left(\mathbf{0},\ \boldsymbol{I}_p\right), & i = 1, \ldots, I,
\end{array}
$$

where $\mathbf{0}$ is a p-vector of zeroes and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. In principle, it is possible to specify priors of distances directly rather than indirectly (i.e., by specifying priors of positions). However, specifying a prior for distances is more challenging than specifying a prior for positions, because the distances must satisfy the triangle inequality. As a consequence, it is conventional in the latent space model literature to place a prior on positions rather than distances, and we follow here convention. We use $\tau_\beta^2 = 4$, $a_\sigma = 1$, $b_\sigma = 1$, $\mu_\gamma = 0.5$, $\tau_\gamma^2 = 1$, which are not uncommon in the literature (see, e.g., Furr et al. 2016,). While these priors may seem strong, note that the effective parameter space of many models for binary item response data is small: For example, when item responses are independent Bernoulli(π) random variables with success probability $\pi \in (0, 1)$ and log odds $\alpha = \text{logit}(\pi) \in \mathbb{R}$, then values of α outside of the interval $[-5, +5]$ correspond to probabilities close to 0 or 1, which are unrealistic. Therefore, while the theoretical parameter space of α is $\mathbb{R}$, the effective parameter space is a subset of $\mathbb{R}$, e.g., $[-5, +5]$. As a consequence, using priors that place most probability mass on $[-5, +5]$ are reasonable, as do the above priors. The priors described above are used throughout the remainder of the paper, unless stated otherwise.

The posterior of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ and the unobserved positions of respondents $\boldsymbol{A}$ and items $\boldsymbol{B}$, given an observation $\boldsymbol{y}$ of $\boldsymbol{Y}$, are proportional to

$$\begin{aligned} f(\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \gamma,\ \boldsymbol{A},\ \boldsymbol{B} \mid \boldsymbol{y}) \propto & \left[\prod_{j=1}^{N} f(\alpha_j)\right] \left[\prod_{i=1}^{I} f(\beta_i)\right] f(\gamma) \left[\prod_{j=1}^{N} f(\boldsymbol{a}_j)\right] \left[\prod_{i=1}^{I} f(\boldsymbol{b}_i)\right] \\ & \times \left[\prod_{j=1}^{N}\prod_{i=1}^{I} \mathbb{P}\Big(Y_{j,i} = y_{j,i} \mid \alpha_j,\ \beta_i,\ \gamma,\ \boldsymbol{a}_j,\ \boldsymbol{b}_i\Big)\right], \end{aligned} \tag{5}$$

where, in an abuse of notation, we use $f(.)$ to denote the prior and posterior probability density functions of the parameters as well as the positions of respondents and items.

We sample from the posterior by using an MCMC algorithm that updates the parameters and the positions of respondents and items at iteration t as follows:

1. Propose $\alpha_j^\star$ from a symmetric proposal distribution and accept the proposal with probability

$$\min\left(1,\ \frac{f\Big(\alpha_j^\star \mid \mathbf{y},\ \boldsymbol{A},\ \boldsymbol{\alpha}_{-j},\ \boldsymbol{B},\ \boldsymbol{\beta},\ \gamma\Big)}{f\Big(\alpha_j^{(t)} \mid \mathbf{y},\ \boldsymbol{A}, \boldsymbol{\alpha}_{-j},\ \boldsymbol{B},\ \boldsymbol{\beta},\ \gamma\Big)}\right),$$

   $\boldsymbol{\alpha}_{-j} = (\alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_N)$
2. Propose $\beta_i^\star$ from a symmetric proposal distribution and accept the proposal with probability

$$\min\left(1,\ \frac{f\Big(\beta_i^\star \mid \mathbf{y},\ \boldsymbol{A},\ \boldsymbol{B}, \boldsymbol{\beta}_{-i},\ \boldsymbol{\alpha},\ \gamma\Big)}{f\Big(\beta_i^{(t)} \mid \mathbf{y},\ \boldsymbol{A},\ \boldsymbol{B}, \boldsymbol{\beta}_{-i},\ \boldsymbol{\alpha},\ \gamma\Big)}\right),$$

   $\boldsymbol{\beta}_{-i}, = (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_I)$
3. Propose $\gamma^\star$ from a symmetric proposal distribution and accept the proposal with probability

$$\min\left(1,\ \frac{f\Big(\gamma^\star \mid \mathbf{y},\ \boldsymbol{A},\ \boldsymbol{B},\ \boldsymbol{\alpha},\ \boldsymbol{\beta}\Big)}{f\Big(\gamma^{(t)} \mid \mathbf{y},\ \boldsymbol{A},\ \boldsymbol{B},\ \boldsymbol{\alpha},\ \boldsymbol{\beta}\Big)}\right),$$

4. Propose $\boldsymbol{a}_j^\star$ from a symmetric proposal distribution and accept the proposal with probability

$$\min\left(1,\ \frac{f\left(\boldsymbol{a}_j^\star \mid \mathbf{y},\ \boldsymbol{A}_{-j},\ \boldsymbol{B},\ \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \gamma\right)}{f\left(\boldsymbol{a}_j^{(t)} \mid \mathbf{y},\ \boldsymbol{A}_{-j},\ \boldsymbol{B},\ \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \gamma\right)}\right),$$

   where $\boldsymbol{A}_{-j} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{j-1}, \boldsymbol{a}_{j+1}, \ldots, \boldsymbol{a}_N)$.
5. Propose $\boldsymbol{b}_i^\star$ from a symmetric proposal distribution and accept the proposal with probability

$$\min\left(1,\ \frac{f\left(\boldsymbol{b}_i^\star \mid \mathbf{y},\ \boldsymbol{A},\ \boldsymbol{B}_{-i},\ \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \gamma\right)}{f\left(\boldsymbol{b}_i^{(t)} \mid \mathbf{y},\ \boldsymbol{A},\ \boldsymbol{B}_{-i},\ \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \gamma\right)}\right),$$

   where $\boldsymbol{B}_{-i} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{i-1}, \boldsymbol{b}_{i+1}, \ldots, \boldsymbol{b}_I)$.
6. Sample σ^2 from its full conditional distribution:

$$\sigma^2 \sim \text{Inv-Gamma}\left(a_\sigma + \frac{N}{2},\ b_\sigma + \frac{\sum_{j=1}^N \alpha_j^2}{2}\right).$$

As symmetric proposal distributions, we use (multivariate) Gaussian distributions centered at the current values of the parameters or the positions of respondents and items, with diagonal variance–covariance matrices. The variances of the (multivariate) Gaussians are set to achieve a good performance of the algorithm (with an acceptance rate of 0.3). To detect non-convergence of the MCMC algorithm, we use trace plots along with the Gelman–Rubin diagnostic (Gelman and Rubin 1992). The MCMC algorithm was written in `R`. The `R` code, along with an example dataset, can be found in the supplementary materials.

3.2. *Identifiability*

The log odds of a correct response is invariant to translations, reflections, and rotations of the positions of respondents and items, because the log odds depends on the positions through the distances, and the distances are invariant under the said transformations. As a consequence, the likelihood function is invariant under the same transformations. The same form of identifiability issue arises in latent space models of network data (Hoff et al. 2002). Such identifiability issues can be resolved by post-processing the MCMC output with Procrustes matching (Gower 1975). However, the results need to be interpreted with care, because there are many latent space configurations that give rise to the same distances. So an estimated latent space should be interpreted in terms of the relative distances between positions, rather than the actual positions.

3.3. *Model Selection*

In practice, given a dataset, it is natural to ask: Did the Rasch model with $\gamma = 0$ or the latent space model with $\gamma > 0$ generate the data? If the latent space model generated the data, it is appropriate to base conclusions regarding respondents and items on the latent space model, including the interaction map provided by the latent space model. Otherwise the Rasch model suffices.

To determine whether the Rasch model with $\gamma = 0$ or the latent space model with $\gamma > 0$ generated the data, we use a model selection approach based on spike-and-slab priors (Ishwaran and Rao 2005). We specify a spike-and-slab prior for $\log \gamma$ by specifying a prior consisting of two component distributions: a spike prior $N_{spike}(\mu_{\gamma_0}, \tau^2_{\gamma_0})$ with a small variance $\tau^2_{\gamma_0} > 0$ that places most of its probability mass in a small neighborhood of 0, and a slab prior $N_{slab}(\mu_{\gamma_1}, \tau^2_{\gamma_1})$ with a large variance $\tau^2_{\gamma_1} > 0$ that distributes its probability mass across the parameter space. In other words, the prior of $\log \gamma$ may be expressed as:

$$\log \gamma \ \sim \ (1-\delta)\ N_{spike}(\mu_{\gamma_0}, \tau^2_{\gamma_0}) + \delta\ N_{slab}(\mu_{\gamma_1}, \tau^2_{\gamma_1}),$$

where $\delta \in \{0, 1\}$. If the posterior probability of the event $\delta = 0$ is less than 0.5, we choose the Rasch model with $\gamma = 0$; otherwise, we choose the latent space model with $\gamma > 0$. The posterior probability can be approximated by the proportion of times we observe the event $\delta = 1$ in a Markov chain Monte Carlo sample from the posterior. We choose as a prior for $\omega = p(\delta = 1 \mid \omega) \in [0, 1]$ the Beta(1, 1) distribution. As a spike prior, we use $N_{spike}(-3, 1)$, so that the distribution of $\log \gamma \mid \delta = 0$ has mode 0.02, mean 0.08, and standard deviation 0.01. As a slab prior, we use $N_{spike}(0.5, 1)$, with mode 0.61, mean 2.72, and standard deviation 3.56. As pointed out in Sect. 3.1, the effective parameter space of many models for binary item response data is small, so the slab prior is not unreasonable.

The model selection method described is applied to the two real data applications provided in Sect. 4. We evaluate the accuracy of the model selection method via simulations in Sect. 5.

4. Applications

To demonstrate the latent space approach, we provide two empirical examples.

4.1. Example 1: Attitudes to Abortion

4.1.1. Data and Estimation As a first example, we used the attitudes-to-abortion scale that came from Social and community planning research (1987). Seven items were included in the scale, which ask respondents whether abortion should be legal in each of the following seven scenarios:

1. The woman decides on her own that she does not wish to have the child
2. The couple agree that they do not wish to have the child
3. The woman is not married and does not wish to marry the man
4. The couple cannot afford any more children.
5. There is a strong chance of a defect in the baby
6. The woman's health is seriously endangered by the pregnancy
7. The woman became pregnant as a result of rape

Binary responses to the seven items were collected, where response 'Yes' was coded as 1 and response 'No' was coded as 0. The mean proportion of 'Yes' was 0.42, 0.52, 0.47, and 0.53 for Items 1 to 4, while 0.86, 0.94, and 0.93 for Items 5 to 7, respectively (N=642).[1] The positive response proportion was quite high for the last three items that describe rather extreme situations in which most respondents are likely to endorse. To implement MCMC, we specified the priors as we described in Sect. 3. For β, we chose a stronger prior with $\tau^2_\beta = 1$ because otherwise the

[1] Different subsets or versions of the data have been used in the literature. We used the data pre-processed by Skrondal and Rabe-Hesketh (2004), which include responses from 734 respondents. We analyzed the version after deleting respondents with no item responses assuming missing at random (Skrondal and Rabe-Hesketh 2004).

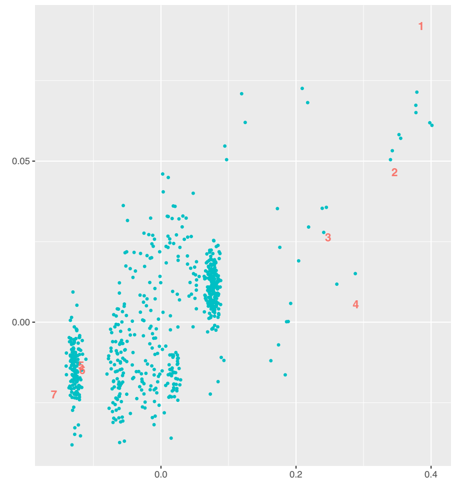

FIGURE 3.
Estimated latent space for the attitudes to abortion data. Red numbers represent items, and blue dots represent respondents

MCMC did not converge well due to the boundary effects of probability for Items 5–7 (the positive answer probability was too close to 1). We selected the tuning parameters (standard deviations of the proposal distributions) to ensure a reasonable acceptance rate as follows: 2.2 for α_j, 0.5 for β_i, 0.1 for γ, 1.7 for $\boldsymbol{a}_j$, and 0.4 for $\boldsymbol{b}_i$. The MCMC run included 20,000 iterations with the first 10,000 iterations discarded as a burn-in period. The computation took approximately 28 minutes for the latent space model on a standard computer. Trace plots showed reasonable convergence of the sampler (convergence evidence was provided in "Appendix B" of the supplement). In addition, we used the Gelman–Rubin diagnostic (Gelman and Rubin 1992) to detect possible non-convergence. We ran the model with three sets of random starting values; the scale reduction factor was smaller than 1.06 for all model parameters, suggesting that there are no signs of non-convergence. We implemented the model selection method with the spike-and-slab prior, described in Sect. 3.3. The posterior inclusion probability of δ was .99, in favor of the proposed model to the Rasch model. Hence, we move forward with the latent space item response model for the current application.

4.1.2. Results

Interpreting latent space results

Figure 3 displays the estimated latent space. This latent space shows the point estimates (posterior means) of the positions but not their uncertainty, for the ease of visualization. Uncertainty of the estimated positions, measured with the 95% posterior credible intervals, is reported in "Appendix C" of the supplement. The γ parameter was estimated as 1.25 (posterior median, with 95% posterior credible interval [0.92, 1.54]) and σ as 2.34 (posterior median, with 95% posterior credible interval [2.07, 2.62]).

Roughly two item groups appear in the latent space: one group with three items in the bottom left of the space (Items 5–7) and the other group with four items in the top right side of the space (Items 1–4). Also, two respondent groups appear, with a larger group in the left bottom part of the space (near Items 5–7) and a much smaller, scattered group on the right upper side of the space (near Items 1–4). Respondents close to Items 1–4 but apart from Items 5–7 tend to give positive responses to the mild items (Items 1–4) but negative responses to the extreme items (Items 5–7). Table 2 shows that the respondents in the region of $X > 0.3$ and $Y > 0.025$ (close to Items 1-4)

TABLE 2.
Response patterns to Items 1 to 7 (I1 to I7) for respondents (ID) in the smaller person cluster, located in the bottom left corner of the latent space (which is the region of $X > 0.3$ and $Y > 0.025$)

ID	I1	I2	I3	I4	I5	I6	I7
27	1	0	0	1	0	0	0
92	1	1	1	1	1	0	0
132	1	1	0	1	0	0	0
191	1	1	1	1	0	1	0
273	1	1	0	0	0	0	0
330	1	0	0	1	0	0	0
653	1	1	0	1	0	0	0
662	1	1	1	0	0	0	0
675	1	1	1	0	0	0	0

These people tend to give positive responses to I1–I4, but negative responses to I5–I7

indeed tend to choose YES to Items 1 to 4 but NO to Items 5 to 7. Respondents close to Items 5–7 tend to give positive responses to the extreme items but negative responses to the mild situations (Items 1–4).

Comparison with the Rasch model

We compared our parameter estimates with those from the Rasch model (Eq. 1). The Rasch model was estimated with the fully Bayesian approach with the same set of priors as our model's for the $\beta_i, \alpha_j, \sigma^2$ parameters. Estimation details were provided in "Appendix D" of the supplement.[2]

From the Rasch model, σ was estimated as 2.40 (posterior median, with 95% posterior credible interval [1.87, 3.02]), similar to the latent space IRT model estimate (2.34 with [2.07, 2.62]). The item parameter estimates (β_i) are displayed in Fig. 4a, b. The item parameter estimates from the Rasch model appear smaller by a constant (approximately 2 in the logit scale) compared with the latent space model, which is sensible given that the proposed model has the additional penalty term (distances).

We then compared the person parameter estimates (α_j) in Fig. 4c, d. Overall, the estimates under these two models are similar, with a rank order correlation of .95. It is noteworthy that α_j rarely change much compared to the Rasch model, whereas β_i alters due to the added latent space term. This is another evidence that the latent space is not the ability space, at least in this example.

Posterior predictive checking

We evaluated the absolute goodness-of-fit of the proposed latent space model based on posterior predictive checking. We compared the proportions of correct responses between the observed data and the replicated datasets (based on the estimated model parameters). Little discrepancy between the observed and replicated measures indicates satisfactory goodness-of-fit of the model. Figures 5 displays the box plots of the predicted proportions of positive responses for the seven items over 10,000 replicated data. The red dot in each box indicates the observed proportion. The predicted measures show highly congruent behavior to the observed measure, suggesting reasonable goodness-of-fit of the proposed model to the data under investigation. Further, based on Cohen's d effect size, no item showed large mean differences between the replicated and original data with the proposed model ($|d| > 0.8$).

[2]The MCMC estimates of the Rasch model were very similar to the ML estimates obtained from the R lme4 package (Bates et al. 2015). The results are also shown in "Appendix D" of the supplement.

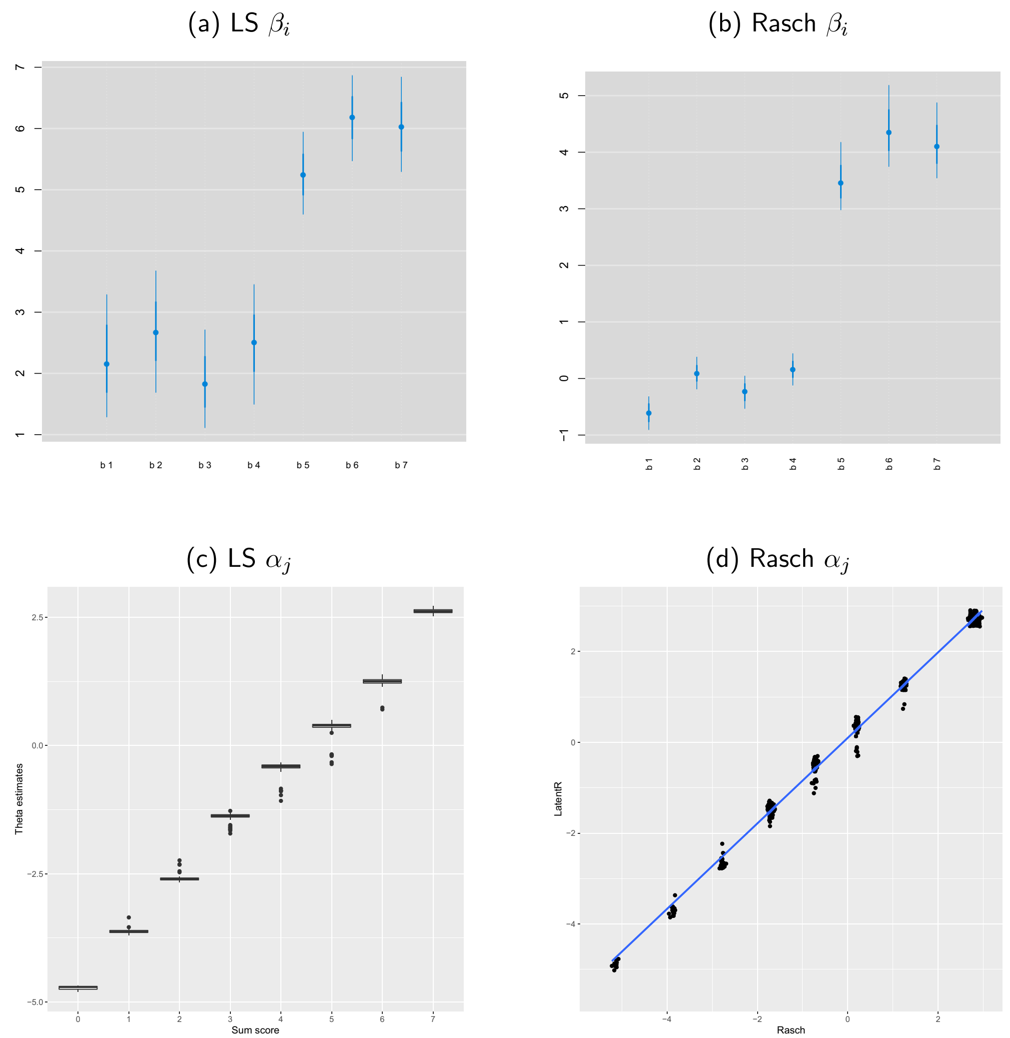

FIGURE 4.
a, **b** The 95% posterior credible intervals for β for Items 1 to 7 (indicated by b1 to b7) from the latent space model (LS) and the Rasch model for the attitudes to abortion data. **c** The box plot of α estimates per total sum score on the X-axis (0 to 7) from the latent space model. **d** α estimates between the Rasch model (Rasch, X-axis) and the latent space model (latent, Y-axis)

4.2. Example 2: Deductive Reasoning

4.2.1. Data and Estimation As a second example, we used the data from the Competence Profile Test of Deductive Reasoning—Verbal assessment (DRV; Spiel et al. 2001; Spiel and Gluck 2008). This dataset was analyzed in Jin and Jeon (2019), allowing us to compare ours to the results from the NIRM approach. The DRV test was developed to measure deductive reasoning of children in different developmental stages and includes 24 binary items (0 = correct, 1 = incorrect), which fall into three broad categories: (1) Type of inference [four levels: Modus Ponens (MP), Modus Tollens (MT), Negation of Antecedent (NA), and Affirmation of Consequence (AC)]; (2) Content of conditional [three levels: Concrete (CO), Abstract (AB), and Counterfactual (CF)]; and (3)

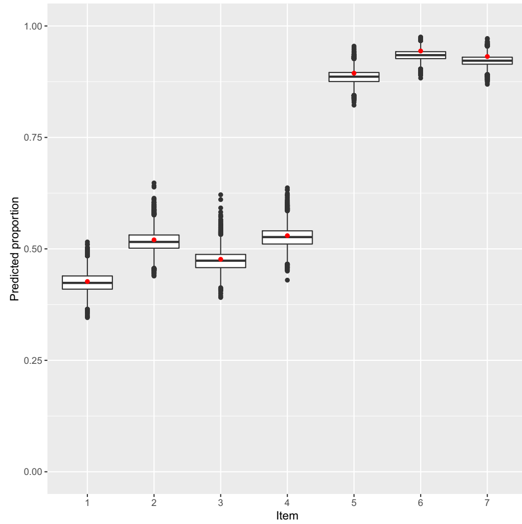

FIGURE 5.
Predicted proportions of the positive responses for the seven items for the attitudes to abortion data. The red dot in each box indicates the proportion of positive responses calculated from the raw data

Precedent of antecedent [two levels: No Negation (UN) and Negation (N)]. More details are provided in "Appendix E" of the supplement.

The data include item responses from 418 school students, 162 female and 256 male students from grade 7 to 12. The success rate ranged from 0.19 to 0.85 with a mean of 0.53 for the 24 test items. The MCMC algorithm described in Sect. 3 was used to sample from the posterior. The standard deviations of the proposal distributions were selected to ensure a reasonable acceptance rate as follows: 0.4 for β, 1.4 for α, 0.05 for γ, 1.1 for a and 0.4 for b. We generated 20,000 MCMC iterations, discarding the first 10,000 iterations as burn-in. The computing time was about 56 minutes. The trace plots do not show obvious signs of non-convergence ("Appendix F" of the supplement). As an additional guard against non-convergence, we used the Gelman–Rubin diagnostic. To do so, we ran the MCMC algorithm with three sets of starting values chosen at random. We found that the scale reduction factor was less than 1.1 for all parameters, so there were no signs of non-convergence. The model selection method described in Sect. 3.3 was used to determine whether the dataset was generated by the Rasch model with $\gamma = 0$ or the latent space model with $\gamma > 0$. A posterior probability of more than .99 in favor of the latent space model suggests that the latent space model generated the data, so all the following results are based on the latent space model.

4.2.2. Results

Parameter estimates

Figure 6a, b shows the 95% posterior credible intervals for the β_i estimates of the 24 items and the distribution of the α_j estimates per total test score. The β_i estimates ranged from 1 to 7 and the α_j estimates ranged from -2 to 2. The α_j estimates were generally aliened well with the total scores. The latent position estimates, posterior means and 95% posterior credible intervals, are provided in "Appendix G" of the supplement. The γ parameter was estimated as 2.23 (posterior median, with 95% posterior credible interval [2.08, 2.35]) and σ as 2.51 [2.27, 2.76].

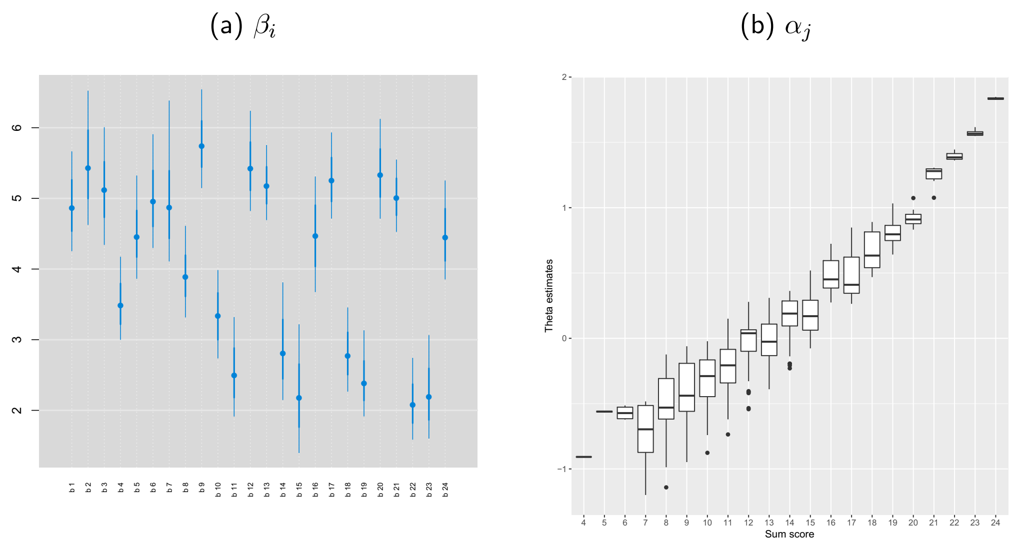

FIGURE 6.
a 95% posterior credible intervals for the β_i estimates (b1 and b24 on the X-axis represent Items 1 to 24), and **b** the distribution of the α_j estimates per total test score for the DRV data. The estimates are from the latent space model

Posterior predictive checking

Goodness-of-fit of the latent space model was evaluated with posterior predictive checking. Figure 7 displays the box plots of the predicted correct response proportions over 10,000 replicated responses for the 24 DRV test items from the proposed model. The red dot in each box indicates the correct response proportion from the original data. The result shows that the prediction of our proposed model was excellent, supporting satisfying goodness-of-fit of the proposed model. Based on Cohen's d effect size, no item showed large mean differences from the original data with the proposed model ($|d| > 0.8$).

Item structure

Figure 8a displays the estimated latent space, where bullet points represent respondents and numbers represent items. Roughly, four item groups appear as color-coded for distinction. The four item group members are listed in Table 3. The item structure identified here shows an excellent agreement with the structure identified in Jin and Jeon (2019) based on the NIRM approach.

I1 and I2 in the upper part of the latent space consist of Concrete items (CO). They are further differentiated in terms of Type of Inference; I1 on the left includes bi-conditional inference items (MP and MT) and I2 on the right includes more complex inference type items (NA and AC). I3 and I4 in the bottom part of the latent space consist of logical fallacy items (Ab and CF). They are further separated by Type of Inference; I3 on the left includes bi-conditional items (MP and MT) and I4 on the right includes complex algebra items (NA and AC). The Presentation of Antecedent factor (UN vs. N) is mixed in all groups, meaning that this factor hardly contributes to item differentiation.

Success probabilities for item group

We assessed the correct response probabilities within and between the four identified item groups (I1, I2, I3, and I4). The density plots of the log odds success probabilities of the individual items per item group are presented in Figure 9. While the four groups show different patterns of the logit success probabilities, items within the same group are similar in the patterns.

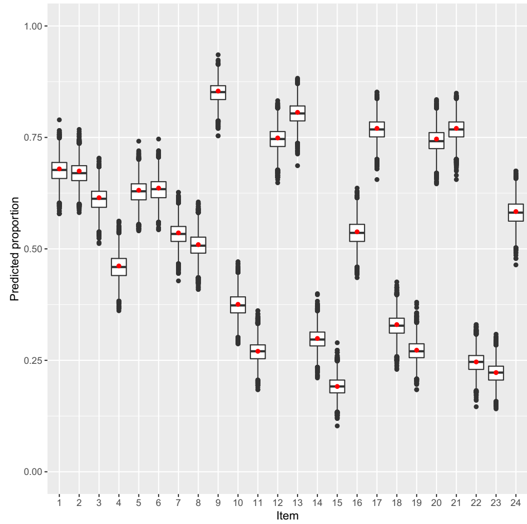

FIGURE 7.
Box plots of the predicted proportions of the correct responses for the 24 DRV test items from 10,000 replicated data. The red dot in each box indicates the proportion of the correct responses for the corresponding item from the raw data (Color figure online)

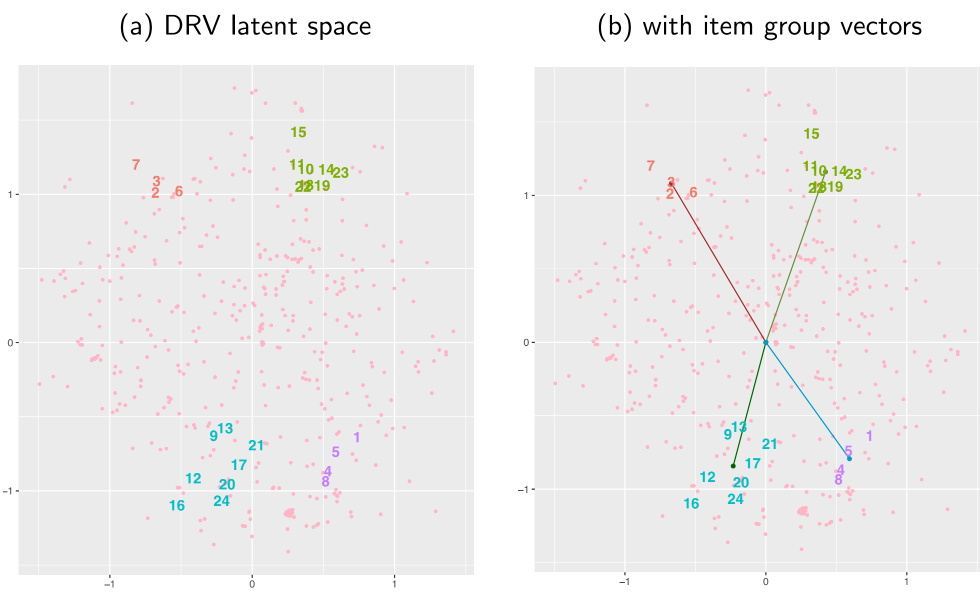

FIGURE 8.
a Latent space for the DRV data and **b** DRV data latent space superimposed with the four vectors that represent the centers of four item groups (I1, I2, I3, and I4). In both figures, dots represent respondents and numbers represent items. Four item groups are distinguished with four different colors. I1: Items 2,3,6,7; I2: Items 10,11,14,15,18,19,22,23; I3: Items 9,12,13,16,17,20,21,24; I4: Items 1,4,5,8

TABLE 3.
Members of the four item groups identified in the DRV data latent space

Item group	Group details
I1	UN_CO_NA (2); UN_CO_AC (3); N_CO_NA (6); N_CO_AC (7)
I2	UN_AB_NA (10); UN_AB_AC (11); N_AB_NC (14); N_AB_AC (15); UN_CF_NA (18); UN_CF_AC (19); N_CF_NA (22); N_CF_MT (23)
I3	UN_AB_MP (9); UN_AB_MT (12); N_AB_MP (13);N_AB_MT (16); UN_CF_MP (17); UN_CF_MT (20); N_CF_MP (21); N_CF_MT (24)
I4	UN_CO_MP (1); UN_CO_MT (4); N_CO_MP (5); N_CO_MT (8)

Numbers in parenthesis indicate item numbers. The acronyms in the item labels indicate the following design factors and their levels: (1) UN vs. N: no negation (UN) and Negation (N) for the presentation of the antecedent factor. (2) CO vs. AB vs. AC: Concrete (CO), Abstract (AB), and Counterfactual (CF) for the content of conditional factor. (3) MP vs. MT vs. NA vs. AC: Modus Ponens (MP), Modus Tollens (MT), Negation of Antecedent (NA), and Affirmation of Consequent (AC) for the type of inference factor

TABLE 4.
Cosine similarity measures between (centers of) the four item groups

	I1	I2	I3	I4
I1	–			
I2	0.618	–		
I3	−0.680	−0.996	–	
I4	−0.996	−0.546	0.613	–

I1: Items 2, 3, 6, 7; I2: Items 10, 11, 14, 15, 18, 19, 22, 23; I3: Items 9, 12, 13, 16, 17, 20, 21, 24; I4: Items 1, 4, 5, 8

Cosine similarity between item groups

We evaluated similarities between two positions in a latent space by the cosine similarity measure. The cosine similarity of two vectors $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^p$ of length $||\boldsymbol{a}||_2 > 0$ and $||\boldsymbol{b}||_2 > 0$ was computed as

$$\cos(\theta) = \frac{\boldsymbol{a}^\top \boldsymbol{b}}{||\boldsymbol{a}||_2 \; ||\boldsymbol{b}||_2},$$

where θ is the angle between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The cosine similarity measure takes on values in the interval $[-1, 1]$. Two vectors pointing into the same direction have a cosine similarity of 1, two vectors with opposite directions have a cosine similarity of -1, and two orthogonal vectors have a cosine similarity of 0.

While cosine similarity can be computed between any two positions in a latent space—including positions of items, respondents, and both items and respondents—we focus here on similarities between the four item groups. Figure 8b shows the original DRV latent space added with the four vectors indicating the centers of the four item groups (where the centers are the mean positions of the corresponding item group members). Table 4 presents a matrix of cosine similarity measures between the four item groups.

Table 4 confirms that I1 is most dissimilar to I4 and I2 is most dissimilar to I3. Marked dissimilarities between I1 and I4 and between I2 and I3 support our earlier finding that Type of Inference (MP/MT vs. NA/AC) most substantially differentiates the DRV test items.

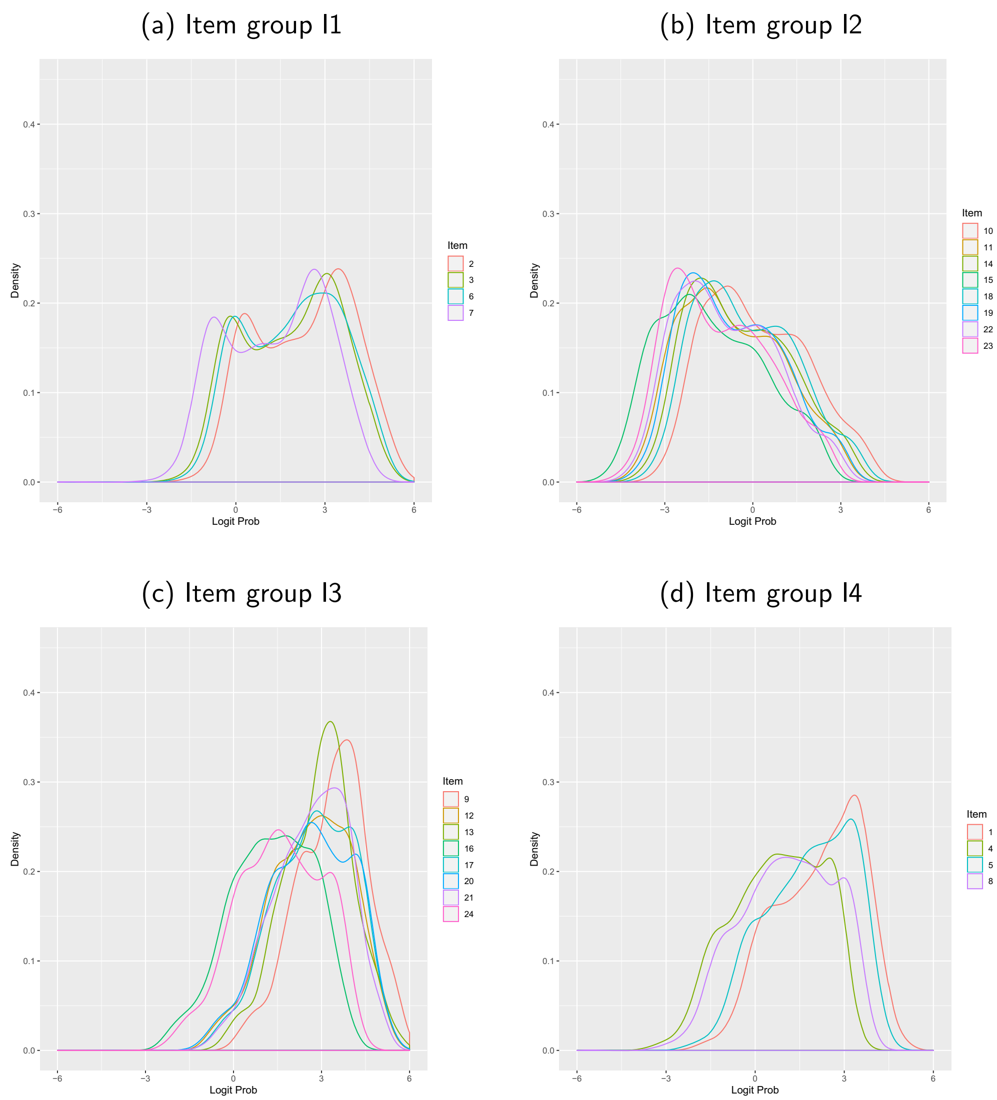

FIGURE 9.
Density plots of the log odds success probabilities of the individual DRV test items per item group. **a**–**d** represent the density plots for item groups I1 to I4 groups in order. I2: Items 10, 11, 14, 15, 18, 19, 22, 23; I3: Items 9, 12, 13, 16, 17, 20, 21, 24; I4: Items 1, 4, 5, 8

Respondent structure

To evaluate their performance in the latent space, we first categorized the children into four sub-groups based on their proximity to the four item groups:

(1) Children near I1. They performed well on logical fallacy inference items (NA/AC) but poorly with simpler inference items (MP/NT) when the items involved concrete conditionals (Co);
(2) Children near I2. They performed well on logical fallacy inference items (NA/AC) but poorly with simpler inference items (MP/NT) if the items involved abstract or counter-factual conditionals (NA/AC);

(3) Children near I3. They performed well on simpler inference items (MP/MT) but poorly with logical fallacy inference items (NA/AC) if the items involved abstract or counterfactual conditionals (NA/AC);
(4) Children near I4. They performed well on simpler inference items (MP/MT) but poorly with logical fallacy inference items (NA/AC) if the items involved concrete conditionals (Co);

Based on the above, we can reasonably conclude that children in sub-groups 3 and 4 were at a lower level of deductive reasoning than those children in sub-groups 1 and 2. While performing well with complex inference items, children in sub-groups 1 and 2 showed poor performance on simpler inference items. This indicates that they might be in a transition to a higher developmental stage. Children in a transition stage tend to make mistakes with easier items, for instance, due to over-generalization on simple problems (e.g., Markovits et al. 1998; Draney 2007,)

Further, it is possible to make additional sub-grouping of children. For instance,

(5) Children between I1 and I3. They were good with both simple and logical fallacy inference items when the items were combined with abstract/counterfactual and concrete conditionals, respectively;
(6) Children between I2 and I4. They were good with both simple and logical fallacy inference items when the items were combined with concrete and abstract/counterfactual conditionals, respectively;
(7) Children around the center of the latent space. They performed equally well on most test items.

How can we identify a specific sub-group for each respondent? One could draw a contour that represents a 95% posterior credible region for each child. If the contours of children overlap, the children may form a subgroup of children that are similar. In addition, one can directly calculate a distance between an individual respondent and each item or item group. For instance, suppose respondent A has a distance of 0.5 from I1, 1.5 from I2, 2 from I3, and 3 from I4. That is, the ratio or relative distances is 1:3:4, meaning that respondent A belongs to the subgroup near I1. So, it is possible to categorize respondents based on their quantified relative distances.

Comparisons with principal component analysis and factor analysis

Our model showed that items in the same item group are similar, while items in different groups are distinctive in terms of contents as well as success probabilities. Traditional methods, such as principal component analysis (PCA) and factor analysis (FA), might be used for similar purposes.

To compare, we applied PCA and FA to the DRV data where two principal components and two factors were extracted.[3] The solutions from the two methods are presented in Fig. 10. Items are placed in the two-dimensional spaces that represent the two principal components or factors. The item clusters identified with PCA and FA are roughly similar to our approach. With factor analysis, the membership of a few items, such as Items 6, 15, 17, and 21, was less clear compared with the other approaches. However, two important differences need to be clarified: (1) in the two traditional methods, extracting factors or principal components (dimensions) is often the main interest, whereas it is not the case in our approach; (2) both in PCA and FA, items and respondents cannot be placed in the same space, unlike our approach.

[3] For PCA, a tetrachoric correlation matrix was used as input data with the R psych package (Revelle 2019). For FA, item factor analysis is applied with oblim rotation by using the R mirt package (Chalmers 2012). With both methods, two-dimensional solutions were optimal.

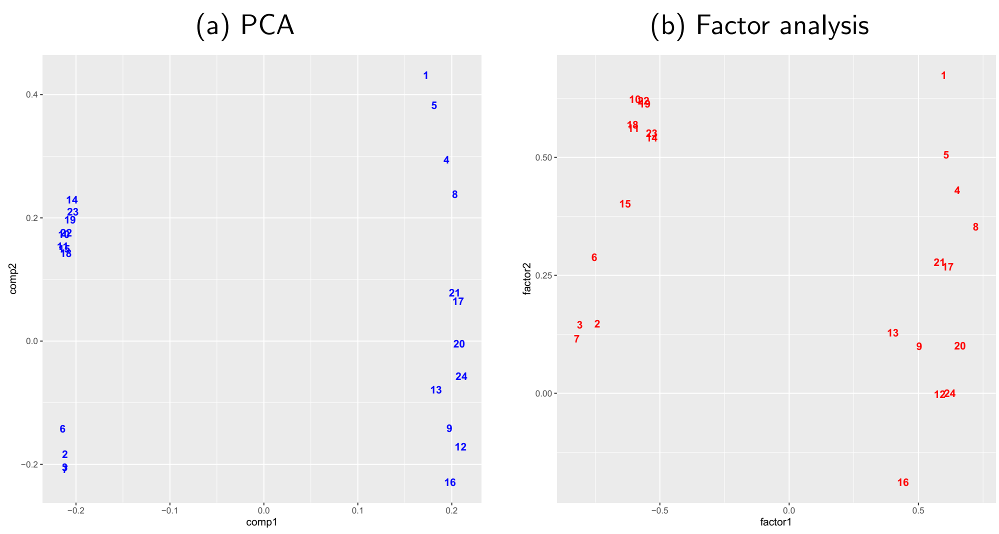

FIGURE 10.
a Principal component analysis (PCA) solution with a tetrachoric correlation matrix as input and **b** item factor analysis (FA) solution with oblim rotation for the DRV data

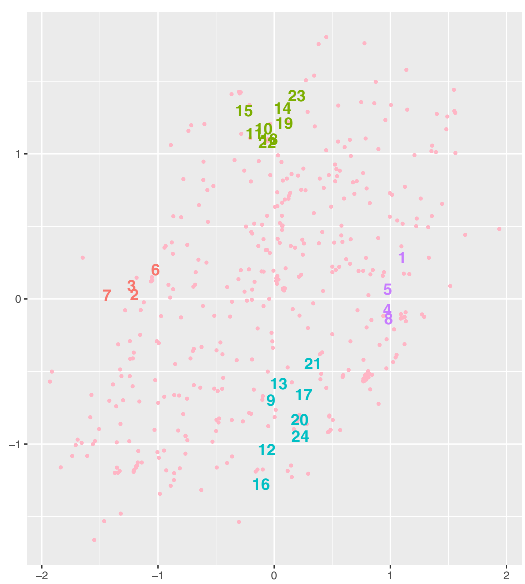

FIGURE 11.
Rotated latent space for the DRV data with oblim rotation. Dots represent respondents and numbers represent items. Four item groups are distinguished with four different colors. I1: Items 2, 3, 6, 7; I2: Items 10, 11, 14, 15, 18, 19, 22, 23; I3: Items 9, 12, 13, 16, 17, 20, 21, 24; I4: Items 1, 4, 5,8 (Color figure online)

Latent space rotation

When desirable, we can attach substantive interpretations to latent space dimensions based on neighboring items. To illustrate, we return to the original latent space displayed in Fig. 8a. No items appear close to the X-axis; therefore, it is difficult to interpret the X-axis in a meaningful way.

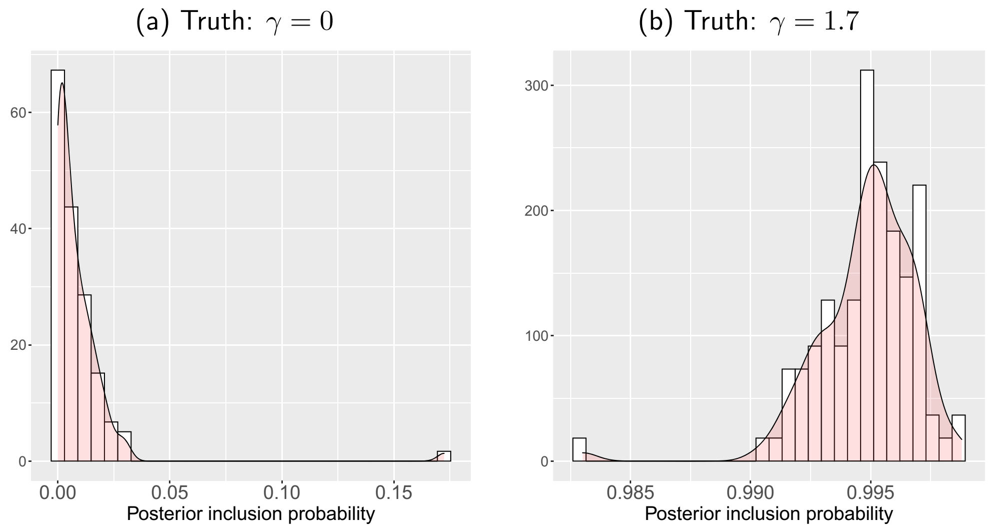

FIGURE 12.
Histogram of the estimated posterior probability of the event $\delta = 1$, called "posterior inclusion probability." **a** Data are generated from the Rasch model with $\gamma = 0$. **b** Data are generated from the latent space model with $\gamma = 1.7$

To improve interpretability, we rotate the original latent space to a place where items are better encompassed by the axes, which is permitted due to rotational invariance property of latent space. Rotation is a frequently utilized technique in factor analysis which also has rotational invariance. We applied oblim rotation (Jennrich 2002) to the estimated item position matrix $\boldsymbol{B}$ using the R package GPArotation (Bernaards and Jennrich 2005), and then rotated the respondent position matrix $\boldsymbol{B}$ in the same way with a common rotation matrix. We denote the rotated item and respondent position matrices by $\boldsymbol{A}^*$ and $\boldsymbol{B}^*$, respectively.

Figure 11 displays the rotated latent space for the DRV data. Two item groups I1 and I4 are positioned close to the X-axis, while I2 and I3 are placed close to the Y-axis in the rotated space. This indicates that the X-axis represents Type of Inference (MP/MT vs. NA/AC) combined with Concrete conditionals, while the Y-axis represents Type of Inference (MP/MT vs. NA/AC) combined with Abstract and Counterfactual conditionals. Items are differentiated based on the type of inference in each dimension, while the two dimensions are separated by the content of conditionals (concrete vs. abstract/counterfactual).

5. Simulation Study

We conducted a simulation study to evaluate whether the model selection approach described in Sect. 3.3 can determine if the Rasch model with $\gamma = 0$ or the latent space model with $\gamma > 0$ generated a given dataset. To do so, we used the setting of Fig. 1a, b with $N = 200$ and $I = 14$. We simulated 100 datasets under the Rasch model with $\gamma = 0$ and under the latent space model with $\gamma = 1.7$. For each simulated dataset, the MCMC algorithm was run for 5000 burn-in iterations, followed by 5000 post-burn-in iterations. We estimated the posterior probability of the indicator $\delta = 1$ by the proportion of times $\delta = 1$ in a Markov chain Monte Carlo sample from the posterior.

Figure 12 shows a histogram of the estimated posterior probability of the event $\delta = 1$. In Fig. 12a, data are generated from the Rasch model with $\gamma = 0$. The posterior probability of $\delta = 1$ is smaller than 0.05 in at least 99% of the simulated data sets. In Fig. 12b, data are generated from the latent space model with $\gamma = 1.7$. Here, the posterior inclusion probability of $\delta = 1$ is greater than 0.99 in at least 99% of the simulated data sets. If the Rasch model is chosen when the estimated posterior probability of $\delta = 1$ is less than 0.5 and otherwise the latent space model is chosen, then the data-generating model is selected in all simulated data sets.

These simulation results provide reasonable evidence that the proposed model selection approach helps determine whether the Rasch model with $\gamma = 0$ or the latent space model with $\gamma > 0$ generated a given dataset. In other words, the model selection approach helps decide whether the Rasch model suffices or whether there are systematic deviations from the Rasch model due to unobserved respondent–item interactions, which the latent space model can capture and represent in a low-dimensional space.

6. Discussion

6.1. Summary

We have introduced a novel approach to modeling item response data, capturing deviations from the Rasch model in the form of respondent-item interactions. While the Rasch model is a classic model and may be a natural starting point in practice, many item response datasets can be expected to exhibit deviations from the respondent and item effects of the Rasch model, which implies that there are unexplained interactions between items and respondents. We have presented evidence of interactions in two empirical examples, but in a number of other datasets we tested, we likewise observed that interactions among respondents and items are present and non-negligible. We propose to capture deviations from the Rasch model in the form of respondent–item interactions by embedding both respondents and items in a low-dimensional latent space, which represents interactions between items, between respondents, and between items and respondents that are not explained by the Rasch model. The proposed latent space approach has technical advantages over conventional IRT modeling approaches, because it makes weaker independence and conditional assumptions than conventional approaches, such as the Rasch model. An additional, intriguing advantage is that it produces a geometrical representation of items and respondents that can provide important insights into how respondents perform on test items.

6.2. Some Final Thoughts on Practical Advantages and Possible Applications

We mention here some final thoughts on practical advantages and possible applications of the proposed latent space approach. First and foremost, if the model selection approach described in Sect. 3.3 determines that $\gamma > 0$, then the data exhibit systematic deviations from the main effects of the Rasch model, that is, respondent–item interactions. The estimated latent space supplies an interaction map that represents those deviations in a low-dimensional space, providing diagnostic feedback on items as well as respondents. For example, the estimated latent space may be useful for assessing whether test items are differentiated or grouped together as blueprinted by test developers: For example, the DRV test was developed based on three design factors, and we found that one design factor (the Presentation of Antecedent) barely contributed to item differentiation and could be dropped without much loss.

In addition, the estimated latent space could help detect unintended or undesirable forms of test-taking behavior. For instance, suppose that a computer-based cognitive test with a time constraint is administered (without permission to skip items) and the estimated latent space reveals that a group of respondents is located close to the last test items in the latent space. That may be an unintended consequence of the fact that most test takers ran out of time and did not respond to the last test items, so that the few respondents who did respond to them are close to those items in the latent space. It goes without saying that such conclusions need to be accompanied by additional evidence (e.g., item response times), but the latent space approach can nonetheless be helpful for diagnosing problems in the first place.

Last, but not least, the proposed latent space approach is useful for providing feedback on the test performance of individual test takers or subgroups of test takers. For example, in the DRV example, we have demonstrated that one can identify items that individual test takers may be struggling with. Such information could guide classroom instruction, and help evaluate and improve intervention programs.

Acknowledgments

Ick Hoon Jin was partially supported by the Yonsei University Research Fund 2019-22-0210 and the Basic Science Research Program through the National Research Foundation of Korea (NRF 2020R1A2C1A01009881). Michael Schweinberger was partially supported by NSF award DMS-1812119.

References

Agarwal, D., & Chen, B.-C. (2009). Regression-based latent factor models. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 19-28.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65,* 676–696.

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology, 64*(9), 1089–1108.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289.

Draney, K. (2007). The Saltus model applied to proportional reasoning data. *Journal of Applied Measurement, 8,* 438–455.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50,* 195–212.

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271–288.

Furr, D. C., Lee, S.-Y., Lee, J.-H., & Rabe-Hesketh, S. (2016). Two-parameter logistic item response model - STAN. https://mc-stan.org/users/documentation/case-studies/tutorialtwopl.html.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Iterative simulation using multiple sequences. *Statistical Science, 7,* 457–472.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika, 40,* 33–51.

Hoff, P. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association,* 286–295.

Hoff, P. (2020). Additive and multiplicative effects network models. *Statistical Science* (to appear).

Hoff, P., Raftery, A., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, 97,* 1090–1098.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics, 33,* 730–773.

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika, 67,* 7–19.

Jin, I. H., & Jeon, M. (2019). A doubly latent space joint model for local item and person dependence in the analysis of item response data. *Psychometrika, 84,* 236–260.

Jin, I. H., Jeon, M., Schweinberger, M., & Lin, L. (2018). Hierarchical network item response modeling for discovering differences between innovation and regular school systems in Korea. Available at. arxiv.org/abs/1810.07876.

Lauritzen, S. (1996). *Graphical models*. Oxford: Oxford University Press.

Markovits, H., Fleury, M.-L., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development, 69,* 742–755.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., & Maris, G. K. J. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research, 53,* 15–35.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. London: Chapman & Hall.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann.
Rasch, G. (1961). On general laws and meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability (volume 4)* (pp. 321–333).
Revelle, W. (2019). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from https://CRAN.R-project.org/package=psych(Rpackageversion1.9.12).
Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282.
Schweinberger, M., & Snijders, T. A. B. (2003). Settings in social networks: A measurement model. In R. M. Stolzenberg (Ed.), *Sociological methodology* (Vol. 33, pp. 307–341). Boston & Oxford: Basil Blackwell.
Sewell, D. K., & Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association, 110,* 1646–1657.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
Smith, A. L., Asta, D. M., & Calder, C. A. (2019). The geometry of continuous latent space models for network data. *Statistical Science, 34,* 428–453.
Social and community planning research. (1987). *British social attitude, the 1987 report*. Aldershot: Gower Publishing.
Spiel, C., & Gluck, J. (2008). A model based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects* (pp. 41–60). Gottingen: Hogrefe.
Spiel, C., Gluck, J., & Gossler, H. (2001). Stability and change of unidimensionality: The sample case of deductive reasoning. *Journal of Adolescent Research, 16,* 150–168.
Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.
Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60,* 181–198.

Manuscript Received: 4 APR 2020
Final Version Received: 29 OCT 2020
Accepted: 23 MAR 2021
Published Online Date: 3 MAY 2021