PSYCHOMETRIKA—VOL. 89, NO. 2, 386–410
JUNE 2024
https://doi.org/10.1007/s11336-023-09936-3

WHAT CAN WE LEARN FROM A SEMIPARAMETRIC FACTOR ANALYSIS OF ITEM RESPONSES AND RESPONSE TIME? AN ILLUSTRATION WITH THE PISA 2015 DATA

YANG LIU AND WEIMENG WANG

UNIVERSITY OF MARYLAND

It is widely believed that a joint factor analysis of item responses and response time (RT) may yield more precise ability scores that are conventionally predicted from responses only. For this purpose, a simple-structure factor model is often preferred as it only requires specifying an additional measurement model for item-level RT while leaving the original item response theory (IRT) model for responses intact. The added speed factor indicated by item-level RT correlates with the ability factor in the IRT model, allowing RT data to carry additional information about respondents' ability. However, parametric simple-structure factor models are often restrictive and fit poorly to empirical data, which prompts under-confidence in the suitablity of a simple factor structure. In the present paper, we analyze the 2015 Programme for International Student Assessment mathematics data using a semiparametric simple-structure model. We conclude that a simple factor structure attains a decent fit after further parametric assumptions in the measurement model are sufficiently relaxed. Furthermore, our semiparametric model implies that the association between latent ability and speed/slowness is strong in the population, but the form of association is nonlinear. It follows that scoring based on the fitted model can substantially improve the precision of ability scores.

Key words: factor analysis, item response theory, response time, PISA, cubic splines, copula, penalized maximum likelihood, cross-validation, model fit, local independence, bootstrap.

1. Introduction

Psychometric investigation on cognitive ability and speed has a long and rich history (e.g., Carroll 1993; Gulliksen 1950; Luce 1986; Thorndike et al. 1926). In the 1926 monograph, Thorndike et al. stated that "level", "extent", and "speed" are three distinct aspects in any measure of performance: While both "level" and "extent" are manifested by correctness of answers and thus can be collectively translated to ability in modern terminology, "the speed of producing any given product is defined, of course, by the time required" (Thorndike et al. 1926, p. 26). The prevalence of computerized test administration and data collection in recent years facilitates the acquisition of response-time (RT) data at the level of individual test items. In parallel, we witnessed a mushrooming development of psychometric models for item responses and RT over the past few decades (see De Boeck & Jeon, 2019; Goldhammer, 2015, for reviews), which in turn gave rise to broader investigations on the relationship between response speed and accuracy in various substantive domains (see Lee & Chen, 2011; Kyllonen & Zu, 2016; von Davier et al., 2019, for reviews). Empirical findings suggested that response speed not only composes proficiency or informs the construct to be measured but also bespeaks secondary test-taking behaviors such as rapid guessing (Deribo et al., 2021; Wise, 2017) , using preknowledge (Qian et al., 2016; Sinharay, 2020; Sinharay & Johnson, 2020) , lacking motivation (Finn, 2015; Thurstone, 1937; Wise & Kong, 2005) , etc.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-023-09936-3.

The manuscript was handled by the ARCS Editor Dr. Nidhi Kohli

Correspondence should be made to Yang Liu, Department of Human Development and Quantitative Methodology, University of Maryland, 3304R Benjamin Bldg, 3942 Campus Dr, College Park, MD 20742, USA. Email: yliu87@umd.edu

Characterizing individual differences in ability and speed with item responses and RT data is in essence a factor analysis problem (Molenaar et al., 2015a,b). The two-factor simple-structure model proposed by van der Linden (2007) was arguably the most popular modeling option so far: Item responses and log-transformed RT variables are treated as two independent clusters of observed indicators for the ability and speed/slowness factors, respectively, and the two latent factors jointly follow a bivariate normal distribution (see Fig. 2 of Molenaar et al., 2015a, for a path-diagram representation). A notable merit of the simple-structure factor model is its plug-and-play nature: Analysts can separately apply standard item response theory (IRT) models for discrete responses (e.g., one-, two-, three-, or four- parameter logistic [1-4PL] model; Birnbaum, 1968; Barton & Lord, 1981) and standard factor analysis models for the continuous log-RT variables (e.g., linear-normal factor model; Jöreskog, 1969), and then simply let the two latent factors covary. Despite its succinctness and popularity, the simple-structure model may fit poorly to empirical data. A highly endorsed interpretation for the lack of fit is that the two inter-dependent latent factors cannot fully explain the dependencies among item-level responses and RT variables. Based on this rationale, numerous diagnostics for residual dependencies and remedial modifications of the simple-structure model have been proposed in the recent literature (e.g., Bolsinova et al., 2017; Bolsinova & Maris, 2016; Bolsinova & Molenaar, 2018; Bolsinova et al., 2017; Bolsinova & Tijmstra, 2016; Glas & van der Linden, 2010; Meng et al., 2015; Ranger & Ortner, 2012; van der Linden & Glas, 2010).

Augmenting standard IRT models with a measurement component for item-level RT may result in more precise ability scores, which is often highlighted as a practical benefit of RT modeling in educational assessment (Bolsinova & Tijmstra, 2018; van der Linden et al., 2010) . Under a simple-structure model with bivariate normal factors, the degree to which item-level RT improves scoring precision is dictated by the strength of the inter-factor correlation (see Study 1 of van der Linden et al., 2010). However, near-zero correlation estimates between ability and speed were sometimes encountered in real-world applications (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2017; Lee & Jia, 2014; van der Linden, Scrams, & Schnipke, 1999). Whenever it happens, analysts are inclined to conclude that item-level RT is not useful for ability estimation at all, or that a less parsimonious factor structure is needed to enhance the utility of RT for scoring purposes (e.g., allowing the log-RT variables to cross-load on the ability factor; Bolsinova & Tijmstra, 2018).

Indeed, van der Linden's (2007) model could be overly restrictive for analyzing item responses and RT data. We, however, do not want to rush to the conclusion that it is the simple factor structure that should be blamed and abandoned. Other parametric assumptions, such as link functions, linear or curvilinear dependencies, and distributions of latent traits and error terms, are also part of the model specification and may contribute to the misfit as well. A fair evaluation on the tenability and usefulness of a simple factor structure demands a version of the model with minimal parametric assumptions other than the simple factor structure itself, which we refer to as a *semiparametric simple-structure model*. Should the semiparametric model still struggle to fit the data adequately, we no longer hesitate to give up on the simple factor structure.

Fortunately, the major components of a semiparametric simple-structure factor analysis have been readily developed in the existing literature. They are

1. A semiparametric (unidimensional) IRT model for dichotomous and polytomous responses (Abrahamowicz & Ramsay, 1992; Rossi et al., 2002) ;
2. A semiparametric (unidimensional) factor model for continuous log-RT variables (Liu & Wang, 2022)
3. A nonparametric copula density estimator for ability and speed/slowness with fixed marginals (Kauermann et al., 2013; Dou et al., 2021) .

As a side remark, we are aware of alternative semiparametric approaches that can be used for each of the above three components: for example, the monotonic polynomial logistic model for item responses (Falk & Cai, 2016a,b), the proportional hazard model (Kang, 2017; Ranger & Kuhn, 2012; Wang et al., 2013b) and the linear transformation model (Wang et al., 2013a) for item-level RT, and the finite normal mixture model (Bauer, 2005; Pek et al., 2009) and the Davidian curve model (Woods & Lin, 2009; Zhang & Davidian, 2001; Zhang et al., 2021) for the joint distribution of latent traits. However, we focus on methods based on smoothing splines in the current analysis. Besides, the simultaneous incorporation of flexible models for all the three components of a simple structure model appears to be novel in the literature of RT modeling. Compared to, e.g., Wang et al. (2013a) and Wang et al. (2013b), in which semiparametric models were applied to only the RT data, our model fares more flexible and thus is more likely to reveal sophisticated dependency patterns in a joint analysis of item responses and RT data.

By retrospectively analyzing a set of mathematics testing data from the 2015 Programme for International Student Assessment (PISA; OECD, 2016), we revisit the following research questions that have only been partially answered previously through parametric simple-structure models:

(1) Is a simple factor structure sufficient for a joint analysis of item response and RT?
(2) How strong are math ability and general processing speed associated in the population of respondents?
(3) To what extent can processing speed improve the precision in ability estimates under a simple-structure model?

It is worth mentioning that the data set was previously analyzed by Zhan et al. (2018) using a variant of van der Linden's (2007) simple-structure model with testlet effects: A higher-order cognitive diagnostics model with testlet effects was used for item responses, a linear-normal factor model was used for log-transformed RT, and the (higher-order) ability and speed factors were assumed to be bivariate normal. Zhan et al. (2018) reported an estimated inter-factor correlation of -0.2 and hence concluded that the association between speed and ability is weak. We are particularly interested in whether their conclusion stands after abandoning inessential parametric assumptions other than the simple factor structure.

The rest of the paper is organized as follows. We first provide a technical introduction of the proposed semiparametric procedure in Sect. 2: The three components of the semiparametric simple-structure model are formulated in Sects. 2.1 and 2.2, penalized maximum likelihood (PML) estimation and empirical selection of penalty weights are outlined in Sect. 2.3, and bootstrap-based goodness-of-fit assessment and inferences are described in Sects. 2.4 and 2.5. Descriptive statistics for the 2015 PISA mathematics data and a plan of our analysis are summarized in Sect. 3, followed by a detailed report of results in Sect. 4. The paper concludes with a discussion of broader implications of our findings and limitations of our method.

2. Methods

2.1. Unidimensional Semiparametric Factor Models

Let $Y_{ij} \in \mathcal{Y}_j \subset \mathbb{R}$ be the jth *manifest variable (MV)* observed for respondent i: Y_{ij} represents either a discrete response to a test item or a continuous item-level RT. In our semiparametric factor model, the distribution of Y_{ij} is characterized by the following logistic conditional density[1] of $Y_{ij} = y \in \mathcal{Y}_j$ given a unidimensional *latent variable* (LV; also known as latent factor, latent trait,

[1] With a slight abuse of terminology, both probability density functions for continuous random variables and probability mass functions for discrete random variables are referred to as densities.

etc.) $X_i = x \in \mathcal{X} \subset \mathbb{R}$:

$$f_j(y|x) = \frac{\exp\big(g_j(x, y)\big)}{\int_{\mathcal{Y}_j} \exp\big(g_j(x, y')\big)\, \mu_j(\mathrm{d}y')}, \tag{1}$$

in which the normalizing integral with respect to the dominating measure μ_j on $\mathcal{Y}_j$ is assumed to be finite. Equation 1 defines a valid conditional density as it is non-negative and integrates to unity with respect to y for a given x. However, the bivariate function $g_j : \mathcal{X} \times \mathcal{Y}_j \to \mathbb{R}$ is not identifiable: It is not difficult to see that adding any univariate function of x to $g_j(x, y)$ does not change the value of Eq. 1 (Gu, 1995, 2013). To impose necessary identification constraints, we re-write g_j by the functional analysis of variance (fANOVA) decomposition

$$g_j(x, y) = g_j^{y}(y) + g_j^{xy}(x, y) \tag{2}$$

and require that

$$g_j^{y}(y_0) = 0,\ g_j^{xy}(x_0, y) \equiv 0,\ \text{and } g_j^{xy}(x, y_0) \equiv 0 \tag{3}$$

for some reference levels $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}_j$. Equation 3 is referred to as *side conditions*; x_0 and y_0 can be set arbitrarily within the respective domains (see Liu & Wang, 2022, for more detailed comments). The univariate component g_j^{y} and the bivariate component g_j^{xy} are functional parameters to be estimated from observed data.

Let $\boldsymbol{\psi}_j : \mathcal{Y}_j \to \mathbb{R}^{L_j}$ be a collection of L_j basis functions defined on the support of Y_{ij}, and $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^{K}$ be a collection of K basis functions defined on the support of X_i. We proceed to approximate the functional parameters by basis expansion. In particular, we set the univariate component

$$g_j^{y}(y) = \boldsymbol{\psi}_j(y)^{\top} \boldsymbol{\alpha}_j, \tag{4}$$

in which the coefficient vector $\boldsymbol{\alpha}_j \in \mathbb{R}^{L_j}$ satisfies

$$\boldsymbol{\psi}_j(y_0)^{\top} \boldsymbol{\alpha}_j = 0. \tag{5}$$

Similarly, the bivariate component is expressed as

$$g_j^{xy}(x, y) = \boldsymbol{\psi}_j(y)^{\top} \mathbf{B}_j \boldsymbol{\varphi}(x), \tag{6}$$

in which the coefficient matrix $\mathbf{B}_j \in \mathbb{R}^{L_j \times K}$ satisfies

$$\mathbf{B}_j \boldsymbol{\varphi}(x_0) = \mathbf{0} \text{ and } \mathbf{B}_j^{\top} \boldsymbol{\psi}_j(y_0) = \mathbf{0}. \tag{7}$$

The linear constraints imposed for the coefficients $\boldsymbol{\alpha}_j$ and $\mathbf{B}_j$ (Eqs. 5 and 7) guarantee that the side conditions (Eq. 3) are satisfied.

Continuous Data When both $X_i \in \mathcal{X}$ and $Y_{ij} \in \mathcal{Y}_j$ (equipped with the Lebesgue measure μ_j) are continuous random variables defined on closed intervals, Eq. 1 corresponds to the

semiparametric factor model considered by Liu and Wang (2022). Without loss of generality, let $\mathcal{X} = \mathcal{Y}_j = [0, 1]$. In fact, any closed interval can be rescaled to the unit interval via a linear transform: If $z \in [a, b]$, $a < b$, then $(z - a)/(b - a) \in [0, 1]$. To approximate smooth functional parameters supported on unit intervals or squares, we use the same cubic B-spline basis with equally spaced knots (De Boor, 1978) for both $\boldsymbol{\psi}_j$ and $\boldsymbol{\varphi}$ (and thus $L_j = K$). It is sometimes desirable to force the MV to be stochastically increasing as the LV increases. Liu and Wang (2022) considered a simple approach to impose likelihood-ratio monotonicity, which boils down to the following linear inequality constraints on the coefficient matrix $\mathbf{B}_j$:

$$(\mathbf{D}_K \otimes \mathbf{D}_K)\,\mathrm{vec}(\mathbf{B}_j) \geq \mathbf{0}. \tag{8}$$

In Eq. 8, $\mathrm{vec}(\cdot)$ denotes the vectorization operator, and

$$\mathbf{D}_K = \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{bmatrix}$$

is a $(K - 1) \times K$ first-order difference matrix. We also set $\mathbf{D}_1 = 1$ by convention.

Discrete Data When $\mathcal{Y}_j = \{0, \ldots, C_j - 1\}$ and μ_j is the associated counting measure, let $y_0 = 0$, $L_j = C_j - 1$, and $\boldsymbol{\psi}_j(y) = (\psi_{j1}(y), \ldots, \psi_{j,C_j-1}(y))^\top$ such that $\psi_{jk}(y) = 1$ if $y = k$ and 0 if $y \neq k$. Then our generic model (Eqs. 1 and 2) reduces to Abrahamowicz and Ramsay's (1992) multi-categorical semiparametric IRT model for unordered polytomous responses, which is further equivalent to the semiparametric logistic IRT proposed by Ramsay and Winsberg (1991) and Rossi et al. (2002) when $C_j = 2$ (i.e., dichotomous data). It is because the basis expansions (i.e., Eqs. 4 and 6) are simplified to $g_j^y(y) = \alpha_{jy}$ and $g_j^{xy}(x, y) = \boldsymbol{\varphi}(x)^\top \boldsymbol{\beta}_{jy}$, in which $\boldsymbol{\beta}_{jy}^\top$ denotes the yth row of $\mathbf{B}_j$, if $y = 1, \ldots, C_j - 1$; meanwhile, $g_j^y(0) = 0$ and $g_j^{xy}(x, 0) \equiv 0$ as part of the side conditions. The conditional density (e.g., Eq. 1) then becomes the item response function (IRF)

$$f_j(y|x) = \begin{cases} \dfrac{1}{1 + \sum_{c=1}^{C_j-1} \exp\left(\alpha_{jc} + \boldsymbol{\varphi}(x)^\top \boldsymbol{\beta}_{jc}\right)}, & \text{if } y = 0, \\ \dfrac{\exp\left(\alpha_{jy} + \boldsymbol{\varphi}(x)^\top \boldsymbol{\beta}_{jy}\right)}{1 + \sum_{c=1}^{C_j-1} \exp\left(\alpha_{jc} + \boldsymbol{\varphi}(x)^\top \boldsymbol{\beta}_{jc}\right)}, & \text{if } y = 1, \ldots, C_j - 1. \end{cases} \tag{9}$$

Like the continuous case, we only consider $\mathcal{X} = [0, 1]$ and $\boldsymbol{\varphi}$ being a cubic B-spline basis defined by a sequence of equally spaced knots. Similar to Eq. 8 in the continuous case, we may impose likelihood-ratio monotonicity on the conditional density by

$$(\mathbf{D}_K \otimes \mathbf{D}_{C_j-1})\,\mathrm{vec}(\mathbf{B}_j) \geq \mathbf{0}, \tag{10}$$

which reduces to $\mathbf{D}_k \boldsymbol{\beta}_{j1} \geq \mathbf{0}$ when $C_j = 2$ (i.e., dichotomous items).

2.2. Simple Factor Structure and Latent Variable Density

Consider a battery of m_1 continuous MVs and m_2 discrete MVs and write $m = m_1 + m_2$. We typically have $m_1 = m_2 = m/2$ when the discrete responses and continuous RT variables are observed for the same set of items. From now on, denote by $Y_{i1}, \ldots, Y_{i,m_1}$ the base-10 log-transformed RT, each of which is rescaled to [0, 1], and by $Y_{i,m_1+1}, \ldots, Y_{im}$ the corresponding responses. Let $X_{i1}, X_{i2} \in [0, 1]$ be the *slowness*[2] and *ability* factors for respondent i, respectively. A *simple factor structure* requires that the item responses $Y_{i,m_1+1}, \ldots, Y_{im}$ are conditionally independent of the slowness factor X_{i1} given the ability factor X_{i2}, and symmetrically that the log-RT variables $Y_{i1}, \ldots, Y_{i,m_1}$ are independent of X_{i2} given X_{i1}. We also make the *local independence* assumption that is standard in factor analysis (McDonald, 1982) : $Y_{i1}, \ldots, Y_{im}$ are mutually independent conditional on X_{i1} and X_{i2}. Further let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^\top$ collect all the MVs produced by respondent i. The simple structure and local independence assumptions imply that

$$f(\mathbf{y}|x_1, x_2) = \prod_{j=1}^{m_1} f(y_j|x_1) \cdot \prod_{j=m_1+1}^{m} f(y_j|x_2), \tag{11}$$

in which $\mathbf{y} = (y_1, \ldots, y_m)^\top \in [0, 1]^{m_1} \times \mathcal{Y}_{m_1+1} \times \cdots \times \mathcal{Y}_m$, and $x_1, x_2 \in [0, 1]$.

For convenience in approximating functional parameters, both X_{i1} and X_{i2} are assumed to follow a Uniform[0, 1] distribution marginally. However, we are aware that uniformly distributed LVs are less attractive for substantive interpretation. Adopting the strategy of Liu and Wang (2022), we define $X^*_{id} = \Phi^{-1}(X_{id})$, $d = 1, 2$, where Φ^{-1} is the standard normal quantile function; the transformed LVs are marginally $\mathcal{N}(0, 1)$ variates, in agreement with the standard formulation in parametric factor analysis. To capture the potentially complex association between latent slowness and ability, we employ a nonparametric estimator for the copula density (Sklar, 1959; Nelsen, 2006) of $(X_{i1}, X_{i2})^\top$, denoted $c(x_1, x_2)$. A copula density is non-negative and has uniform marginals: That is,

$$c(x_1, x_2) \geq 0 \text{ and } \int_0^1 c(x_1, x_2)\mathrm{d}x_1 = \int_0^1 c(x_1, x_2)\mathrm{d}x_2 \equiv 1, \ \forall x_1, x_2 \in [0, 1]. \tag{12}$$

c is in fact the joint density of $(X_{i1}, X_{i2})^\top$ since both X_{i1} and X_{i2} are marginally uniform. In the light of Sklar's theorem, the joint density of the transformed $(X^*_{i1}, X^*_{i2})^\top$ can be calculated by

$$h(x^*_1, x^*_2) = c(\Phi(x^*_1), \Phi(x^*_2))\phi(x^*_1)\phi(x^*_2), \tag{13}$$

in which ϕ and Φ are the density and distribution functions of $\mathcal{N}(0, 1)$, respectively.

We approximate the bivariate copula density c by a tensor-product spline (Dou et al., 2021; Kauermann et al., 2013) :

$$c(x_1, x_2) = \boldsymbol{\varphi}(x_2)^\top \Xi \boldsymbol{\varphi}(x_1) \tag{14}$$

in which $\boldsymbol{\varphi} : [0, 1] \to \mathbb{R}^K$ is a set of cubic B-spline basis functions defined with equally spaced knots[3], and Ξ is an $K \times K$ coefficient matrix. For Eq. 14 to be a proper copula density, we impose the following linear constraints on Ξ:

$$\xi_{kl} \geq 0, \ \forall k, l = 1, \ldots, K, \text{ and } \Xi\boldsymbol{\kappa} = \Xi^\top\boldsymbol{\kappa} = \mathbf{1}, \tag{15}$$

[2]Slowness is the reversal of speed. We abide by the convention that the LV is positively associated with the MV.

[3]For simplicity, the same set of basis functions is used for the LVs in Eqs. 6, 9, and 14.

in which ξ_{kl} is the (k, l)th element of Ξ, and

$$\boldsymbol{\kappa} = \int_0^1 \boldsymbol{\varphi}(x)\mathrm{d}x \tag{16}$$

is a $K \times 1$ vector of normalizing constants for basis functions. It can be verified by elementary properties of B-splines and straightforward algebra that Eqs. 14 and 15 imply Eq. 12.

2.3. Estimation

For each MV $j = 1, \ldots, m$, let $\boldsymbol{\theta}_j = (\boldsymbol{\alpha}_j^\top, \mathrm{vec}(\mathbf{B}_j)^\top)^\top$ collect all the coefficients in g_j^y and g_j^{xy}. Also let $\boldsymbol{\theta} = (\boldsymbol{\theta}_j^\top, \ldots, \boldsymbol{\theta}_m^\top, \mathrm{vec}(\Xi)^\top)^\top$ denote all the coefficients in the simple-structure factor model. We estimate $\boldsymbol{\theta}$ by *penalized maximum (marginal) likelihood (PML)*. The marginal likelihood for the MV vector $\mathbf{Y}_i = \mathbf{y}$ amounts to the integration of Eq. 11 over x_1 and x_2 under the copula density $c(x_1, x_2)$: That is,

$$f(\mathbf{y}; \boldsymbol{\theta}) = \iint_{[0,1]^2} f(\mathbf{y}|x_1, x_2)c(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2. \tag{17}$$

Pooling across an independent and identically distributed (i.i.d.) sample of size n, we arrive at the sample log-likelihood function

$$\ell(\boldsymbol{\theta}; \mathbf{y}_{1:n}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i; \boldsymbol{\theta}), \tag{18}$$

in which $\mathbf{y}_{1:n} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$ denotes an $n \times m$ matrix of observed MV data.

To avoid overfitting, we regularize the roughness of estimated functional parameters by quadratic-form penalties in spline coefficients. For a continuous MV j, the penalty term for $\boldsymbol{\theta}_j$ is the sum of a univariate P-spline penalty for $\boldsymbol{\alpha}_j$ and a bivariate P-spline penalty for $\mathbf{B}_j$ (Eilers & Marx, 1996; Currie et al., 2006) :

$$q_j(\boldsymbol{\theta}_j; \lambda_j) = \frac{\lambda_j}{2}\boldsymbol{\alpha}_j^\top \mathbf{E}_K^\top \mathbf{E}_K \boldsymbol{\alpha}_j + \frac{\lambda_j}{2}\mathrm{vec}(\mathbf{B}_j)^\top \left(\mathbf{I}_K \otimes \mathbf{E}_K^\top \mathbf{E}_K + \mathbf{E}_K^\top \mathbf{E}_K \otimes \mathbf{I}_K\right) \mathrm{vec}(\mathbf{B}_j), \tag{19}$$

in which $\lambda_j > 0$ is the *penalty weight*, $\mathbf{I}_K$ denotes a $K \times K$ identity matrix, and

$$\mathbf{E}_K = \begin{bmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{bmatrix}$$

is a second-order difference matrix of dimension $(K - 2) \times K$. If the MV is polytomous, no penalty is needed for the intercepts $\boldsymbol{\alpha}_j$ and columns of $\mathbf{B}_j$. The resulting P-spline penalty term then becomes

$$q_j(\boldsymbol{\theta}_j; \lambda_j) = \frac{\lambda_j}{2}\mathrm{vec}(\mathbf{B}_j)^\top \left(\mathbf{E}_K^\top \mathbf{E}_K \otimes \mathbf{I}_{C_j - 1}\right) \mathrm{vec}(\mathbf{B}_j). \tag{20}$$

A similar bivariate P-spline penalty is also introduced for the coefficient matrix Ξ:

$$q(\Xi; \lambda_{m+1}) = \frac{\lambda_{m+1}}{2} \mathrm{vec}(\Xi)^\top \left(\mathbf{I}_K \otimes \mathbf{E}_K^\top \mathbf{E}_K + \mathbf{E}_K^\top \mathbf{E}_K \otimes \mathbf{I}_K \right) \mathrm{vec}(\Xi) \tag{21}$$

with a positive penalty weight λ_{m+1}. Combining Eqs. 18–21, we express the penalized sample log-likelihood function as

$$p(\boldsymbol{\theta}; \mathbf{y}_{1:n}, \boldsymbol{\lambda}) = \ell(\boldsymbol{\theta}; \mathbf{y}_{1:n}) - n \left[\sum_{j=1}^{m} q_j(\boldsymbol{\theta}_j, \lambda_j) + q(\Xi; \lambda_{m+1}) \right], \tag{22}$$

in which $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m, \lambda_{m+1})^\top \in (0, \infty)^{m+1}$. PML estimation amounts to finding $\boldsymbol{\theta}$ that maximizes Eq. 22 subject to a series of linear equality and inequality constraints (i.e., Eqs. 5, 7, 8, 10, and 15), which is accomplished by a modified expectation-maximization (EM; Bock & Aitkin, 1981; Dempster et al., 1977) algorithm. A sequential quadratic programming algorithm (Nocedal & Wright, 2006, Algorithm 18.3) is employed in the M-step to handle constrained optimization. The algorithm is a simple extension to what was described in Sects. 4.1 and 4.2 of Liu and Wang (2022); further details are therefore omitted for succinctness. Denote by $\hat{\boldsymbol{\theta}}(\mathbf{y}_{1:n}, \boldsymbol{\lambda})$ the PML estimates of $\boldsymbol{\theta}$ obtained from data $\mathbf{y}_{1:n}$ and penalty weights $\boldsymbol{\lambda}$.

Larger penalty weights enforce less variable yet more biased solutions and *vice versa*—a well-known phenomenon referred to as the *bias-variance trade-off*. To strike a balance, we select the optimal $\boldsymbol{\lambda}$ from a pre-specified grid by multi-fold cross-validation. Let $\Omega_1, \ldots, \Omega_S$ be a partition of the sample: $\bigcup_{s=1}^{S} \Omega_s = \{1, \ldots, n\}$ and $\Omega_s \cap \Omega_{s'} = \emptyset$ for all $s \neq s'$. For each s, let Ω_s^c be the *calibration set* and Ω_s be the *validation set*, in which the superscript c denotes the complement of a set. Predictive adequacy associated with a particular $\boldsymbol{\lambda}$ is gauged by the *empirical risk*

$$R(\mathbf{y}_{1:n}, \boldsymbol{\lambda}) = -\frac{1}{S} \sum_{s=1}^{S} \frac{1}{|\Omega_s|} \ell(\hat{\boldsymbol{\theta}}(\mathbf{y}_{\Omega_s^c}, \boldsymbol{\lambda}); \mathbf{y}_{\Omega_s}), \tag{23}$$

in which $|\Omega_s|$ denotes the size of Ω_s, $\ell(\hat{\boldsymbol{\theta}}(\mathbf{y}_{\Omega_s^c}, \boldsymbol{\lambda}); \mathbf{y}_{\Omega_s})$ denotes the log-likelihood of the validation sub-sample evaluated at the estimated coefficients from the calibration set. Instead of choosing $\boldsymbol{\lambda}$ that minimizes Eq. 23 (i.e., the best solution), we adopt the "one standard error (SE)" heuristic (Chen & Yang, 2021; Hastie et al., 2009) to take into account sampling variability: We select the smoothest solution within one SE from the $\boldsymbol{\lambda}$ that minimizes the empirical risk, where the SE at a specific $\boldsymbol{\lambda}$ is estimated by

$$\mathrm{SE}(\mathbf{y}_{1:n}, \boldsymbol{\lambda}) = \sqrt{\frac{1}{S-1} \sum_{s=1}^{S} \left[-\frac{1}{|\Omega_s|} \ell(\hat{\boldsymbol{\theta}}(\mathbf{y}_{\Omega_s^c}, \boldsymbol{\lambda}); \mathbf{y}_{\Omega_s}) - R(\mathbf{y}_{1:n}, \boldsymbol{\lambda}) \right]^2}. \tag{24}$$

The value $\boldsymbol{\lambda}$ contains $m + 1$ elements. To alleviate the computational burden for penalty weights selection, we set $\lambda_1 = \cdots = \lambda_{m_1} = \lambda_{(c)}$ for continuous MVs, $\lambda_{m_1+1} = \cdots = \lambda_m = \lambda_{(d)}$ for discrete MVs, and $\lambda_{m+1} = \lambda_{(g)}$ for the copula density of the two LVs. We also resort to a multistage workaround to select the remaining three penalty weights: (1) A unidimensional model is fitted to only the continuous MVs to find the optimal $\lambda_{(c)}$, (2) a unidimensional model is fitted to only the discrete MVs to find the optimal $\lambda_{(d)}$, and (3) a two-dimensional simple-structure

model is fitted to all the MVs to find the optimal $\lambda_{(g)}$ while fixing $\lambda_{(c)}$ and $\lambda_{(d)}$ at their optimal values determined in earlier stages. The optimal weights thereby selected are denoted $\hat{\boldsymbol{\lambda}}(\mathbf{y}_{1:n})$. We then refit the model using the optimal weight and the full set of data to obtain the final solution of spline coefficients $\hat{\boldsymbol{\theta}}(\mathbf{y}_{1:n}, \hat{\boldsymbol{\lambda}}(\mathbf{y}_{1:n}))$.

2.4. Model Fit Diagnostics and Inferences

We quantify the sampling variability of sample statistics, including goodness of fit diagnostics and approximations to functional parameters, by bootstrapping (Efron & Tibshirani, 1994; Hastie et al., 2009) . Let $\bar{\mathbf{Y}}_i$ be a random sample from the collection of observed MV vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ such that each element is selected with probability $1/n$. Sample with replacement n times and denote the resulting bootstrap sample $\bar{\mathbf{Y}}_{1:n}$. We approximate the sampling distribution of any test statistic $T(\mathbf{Y}_{1:n})$ by the bootstrap sampling distribution of $T(\bar{\mathbf{Y}}_{1:n})$ conditional on $\mathbf{y}_{1:n}$. Note that most of the test statistics under investigation depend on the optimal penalty weights $\hat{\boldsymbol{\lambda}}(\mathbf{y}_{1:n})$, which is a function of the observed data. Pilot runs suggest that the variability of the optimal weights is small over bootstrap samples; we therefore treat $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}(\mathbf{y}_{1:n})$ as fixed and do not repeat penalty weight selection in the resampling process, which substantially reduces computational time.

Let $S_{ij} = \varsigma_j(Y_{ij})$ be the *MV score* associated with the individual response entry Y_{ij}. For continuous log-RT variables and dichotomous items, we simply let ς_j be the identity function and thus $S_{ij} = Y_{ij}$; for unordered polytomous items, however, a customized ς_j function is needed for recoding raw responses to a more meaningful scale (see Sect. 3.3 for an example). To assess the lack-of-fit for the simple-structure semiparametric model—in particular the unaccounted dependencies residing in observed MVs, we compute the *residual correlation* statistic

$$e_{jj'}(\mathbf{y}_{1:n}, \boldsymbol{\lambda}) = r_{jj'} - \rho_{jj'}(\hat{\boldsymbol{\theta}}(\mathbf{y}_{1:n}, \boldsymbol{\lambda})). \tag{25}$$

for $j, j' = 1, \ldots, m$, $j < j'$. In Eq. 25, $r_{jj'}$ and $\rho_{jj'}$ are the respective sample and model-implied correlations between the jth and j'th MV scores: The model-implied correlation can be further expressed as

$$\rho_{jj'} = \frac{\mu_{jj'} - \mu_j \mu_{j'}}{\sqrt{(\mu_{jj} - \mu_j^2)(\mu_{j'j'} - \mu_{j'}^2)}}, \tag{26}$$

in which we drop the dependency on $\boldsymbol{\theta}$ for conciseness. In Eq. 26, the first moment μ_j can be computed as

$$\mu_j = \int_{\mathcal{Y}_j} \varsigma_j(y) \left[\int_0^1 f_j(y|x) \mathrm{d}x \right] \mathrm{d}y. \tag{27}$$

There are three cases when computing the second moment $\mu_{jj'}$: (1) for a single MV, i.e., $j = j'$,

$$\mu_{jj} = \int_{\mathcal{Y}_j} \varsigma_j(y)^2 \left[\int_0^1 f_j(y|x) \mathrm{d}x \right] \mathrm{d}y; \tag{28}$$

(2) when $j \neq j'$ but the two MVs load on the same LV,

$$\mu_{jj'} = \int_{\mathcal{Y}_j \times \mathcal{Y}_{j'}} \varsigma_j(y) \varsigma_{j'}(z) \left[\int_0^1 f_j(y|x) f_{j'}(z|x) \mathrm{d}x \right] \mathrm{d}y \mathrm{d}z; \tag{29}$$

and (3) when the jth and j'th MVs load respectively on the first and second LVs,

$$\mu_{jj'} = \int_{\mathcal{Y}_j \times \mathcal{Y}_{j'}} \varsigma_j(y)\varsigma_{j'}(z) \left[\int_{[0,1]^2} f_j(y|x_1) f_{j'}(z|x_2) c(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2 \right] \mathrm{d}y \mathrm{d}z. \tag{30}$$

2.5. *Latent Variable Density and Scores*

As we have mentioned in Sect. 2.2, inferences for LVs are made based on the marginally normal X^*_{i1} and X^*_{i2}. In particular, we are interested in the strength of association between the two LVs. To this end, we compute the coefficient of determination for predicting ability (X^*_{i2}) by slowness (X^*_{i1}):

$$\begin{aligned} \eta^2 =& 1 - \frac{\mathbb{E}\left[\mathrm{Var}(X^*_{i2}|X^*_{i1})\right]}{\mathrm{Var}(X^*_{i2})} = \frac{\mathrm{Var}\left[\mathbb{E}(X^*_{i2}|X^*_{i1})\right]}{\mathrm{Var}(X^*_{i2})} \\ =& \int_{\mathbb{R}} \left[\int_{\mathbb{R}} x^*_2 h(x^*_2|x^*_1) \mathrm{d}x^*_2 \right]^2 h(x^*_1) \mathrm{d}x^*_1, \end{aligned} \tag{31}$$

in which $h(x^*_d) = \int_{\mathbb{R}} h(x^*_1, x^*_2) \mathrm{d}x^*_{3-d}$, $d = 1, 2$, is the marginal density of X^*_{id} (assumed to be standard normal), and $h(x^*_2|x^*_1) = h(x^*_1, x^*_2)/h(x^*_1)$ is the conditional density of x^*_2 given x^*_1. Equation 31 reduces to the usual coefficient of determination for linear models when $(X^*_{i1}, X^*_{i2})^\top$ follows a bivariate normal distribution. When analyzing real data, we evaluate η^2 using the estimated LV density, denoted $\hat{\eta}^2(\mathbf{y}_{1:n}, \boldsymbol{\lambda})$; the sampling variability of $\hat{\eta}^2(\mathbf{y}_{1:n}, \boldsymbol{\lambda})$ is again characterized by bootstrapping (Sect. 2.4).

For each respondent i, LV scores can be predicted based on the posterior distribution of $(X^*_{i1}, X^*_{i2})^\top$ given $\mathbf{y}_i = \mathbf{y}$ with density

$$f^h(x^*_1, x^*_2|\mathbf{y}) = \frac{f(\mathbf{y}|\Phi(x^*_1), \Phi(x^*_2))h(x^*_1, x^*_2)}{\iint_{\mathbb{R}^2} f(\mathbf{y}|\Phi(x^*_1), \Phi(x^*_2))h(x^*_1, x^*_2)\mathrm{d}x^*_1 \mathrm{d}x^*_2}. \tag{32}$$

The means of the posterior distribution are often referred to as the expected *a posteriori* (EAP) scores, and the corresponding standard deviations (SDs) gauge the precision of the EAP scores (Thissen & Wainer, 2001) . In practice, density functions involved in Eq. 32 must be estimated from sample data, which introduces additional uncertainty to scores computed from the estimated posterior. Better precision measures can be obtained from a predictive distribution of LV scores (Liu & Yang, 2018a,b; Yang et al., 2012) . Let $\hat{f}^h(x^*_1, x^*_2|\mathbf{y}; \mathbf{y}_{1:n}, \boldsymbol{\lambda})$ be the estimated posterior density. The bootstrap expectation $\mathbb{E}\hat{f}^h(x^*_1, x^*_2|\mathbf{y}; \bar{\mathbf{Y}}_{1:n}, \boldsymbol{\lambda})$ with respect to the (random) bootstrap sample $\bar{\mathbf{Y}}_{1:n}$ defines a suitable predictive density; the inverse variance of the predictive distribution, which is henceforth referred to as the *predictive precision*, can be conveniently estimated from a collection of bootstrap samples. To set the baseline for assessing the gain in predictive precision, we also consider the marginal posterior density of the ability factor X^*_{i2}:

$$f^h(x^*_2|y_{m_1+1}, \ldots, y_m) = \frac{\prod_{j=m_1+1}^m f(y_j|\Phi(x^*_2))h(x^*_2)}{\int_{\mathbb{R}} \prod_{j=m_1+1}^m f(y_j|\Phi(x^*_2))h(x^*_2)\mathrm{d}x^*_2}. \tag{33}$$

Estimated marginal EAP scores and the associated bootstrap predictive precisions can be obtained in a fashion similar to the two-dimensional case.

3. Data and Analysis Plan

3.1. PISA 2015 Mathematics Data

The data we analyze next came from the PISA 2015 computer-based mathematics assessment (OECD, 2016) . The test is composed of 17 dichotomously scored items from two mathematics testing clusters (M1 and M2). Similar to the Zhan et al. (2018), we only retained cases with complete response entries, leading to a total number of $n = 8606$ observations from 58 countries/economies.

Among the 17 items, there are four testlets (with item labels starting with CM155, CM411, CM496, and CM564), each of which involves a pair of items. We collapsed the two items within each testlet into a single four-category nominal item: The four categories 0, 1, 2, and 3 indicated the original item response patterns (0, 0), (1, 0), (0, 1), and (1, 1), respectively. The corresponding RT entries were also summed to a single testlet-level RT variable. Accordingly, the number of items involved in the initial fitting is $m_1 = m_2 = 13$, and the number of MVs is $m = 26$. During data preprocessing, we identified a number of extremely small and large RT entries, which are potential outliers and may cause instability in model fitting. Therefore, we excluded for each MV the top and bottom 1% RT and the associated item response data[4]. Then we took the base-10 logarithm of the RT variables and rescaled them to the unit interval. Selected descriptive statistics of the final data can be found in Tables 1 and 2.

3.2. Analysis Plan

As we have mentioned in Sect. 1, the data set was analyzed in the previous work by Zhan et al. (2018) using a parametric simple-structure model. Though we acknowledge the parsimony and thus retain a simple factor structure, our analysis differs substantially from the previous work, because we model MV-LV and LV-LV dependencies in a nonparametric fashion and are able to provide an ultimate assessment for the validity of a simple factor structure in this data set. Once we confirm that the dependencies in the MVs are sufficiently accounted for, we present graphics and statistics based on the fitted model to demonstrate how the respective distributions of item responses and RT are governed by the ability and slowness factors, as well as how ability and slowness covary in the population of respondents.

Major steps of our analysis are outlined as follows.

step 1. Determine the optimal penalty weights $\hat{\boldsymbol{\lambda}}(\mathbf{y}_{1:n})$ by the three-stage procedure described in Sect. 2.3.
step 2. Draw $B = 100$ bootstrap samples (i.e., resample with replacement) from the observed data $\mathbf{y}_{1:n}$ and repeat model fitting in each bootstrap sample with $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}(\mathbf{y}_{1:n})$.
step 3. Examine the residual correlation statistics (Eq. 25) for all pairs of MVs. Flag a pair if the 90% two-sided bootstrap CI for the residual correlation fall entirely above 0.1 or below −0.1.
step 4. Remove problematic items from the test and repeat steps 1–3 until no large residual correlation remains.
step 5. Plot the conditional densities of the MVs given the marginally normal LVs (Eq. 1 with $x = \Phi(x^*)$) and the joint density of the two LVs (Eq. 13). Compute estimated η^2 statistics (Eq. 31), EAP scores, and the associated predictive SDs for the scores.

Per the request from two referees, we also report in the supplementary document the empirical risk statistics and density estimates for two parametric models. The first model is a standard

[4]Zhan et al. (2018) did not delete any extreme RT entries in their analysis. They performed Bayesian estimation with a somewhat informative prior configuration, which is presumably more stable in the presence of outlying observations.

TABLE 1.
Descriptive statistics for transformed response time (RT).

	CM033Q01	CM474Q01	CM155	CM411	CM803Q01	CM442Q02	
Mean	0.47	0.42	0.71	0.68	0.53	0.64	
SD	0.20	0.19	0.12	0.16	0.18	0.16	
Skew	0.14	0.54	−0.94	−1.09	−0.02	−0.69	
Kurt	2.60	3.08	5.64	5.14	2.89	4.02	
CorrTotal	0.41	0.42	0.50	0.50	0.52	0.58	
	CM034Q01	CM305Q01	CM496	CM423Q01	CM603Q01	CM571Q01	CM564
Mean	0.57	0.59	0.66	0.50	0.69	0.62	0.59
SD	0.18	0.16	0.17	0.17	0.18	0.20	0.16
Skew	−0.28	−0.28	−0.98	0.11	−1.30	−0.92	−0.65
Kurt	3.04	3.46	4.53	3.00	4.84	3.49	4.07
CorrTotal	0.56	0.55	0.45	0.48	0.57	0.52	0.47

For CM155, CM411, CM496, and CM564, summary statistics are computed for the log-transformed RT of the testlets. SD: Standard deviation. Skew: Skewness. Kurt: Kurtosis. CorrTotal: Correlation with the total sum of log-RT.
We applied the based-10 logarithm to the raw RT data and then rescaled the log-transformed variables to [0, 1].

TABLE 2.
Descriptive statistics for item responses.

	CM033Q01	CM474Q01	CM155	CM411	CM803Q01	CM442Q02	
P1	0.77	0.66	0.28	0.21	0.26	0.32	
P2	–	–	0.11	0.19	–	–	
P3	–	–	0.43	0.29	–	–	
CorrTotal	0.44	0.48	0.49	0.55	0.56	0.58	
	CM034Q01	CM305Q01	CM496	CM423Q01	CM603Q01	CM571Q01	CM564
P1	0.38	0.43	0.07	0.79	0.37	0.41	0.22
P2	–	–	0.24	–	–	–	0.19
P3	–	–	0.43	–	–	–	0.27
CorrTotal	0.56	0.31	0.56	0.33	0.45	0.54	0.43

For testlets CM155, CM411, CM496, and CM564, the original item response patterns (0, 0), (1, 0), (0, 1), and (1, 1) are recoded to 0, 1, 2, and 3, respectively. P1–3: Observed proportions of response categories 1–3. CorrTotal: Correlation with total score; we apply the scoring function described in Sect. 3.3 to the testlet MVs before computing the total score.

baseline model for the joint analysis of item response and RT data, which features linear-normal factor models for log-RT variables, 2PL models for item responses, nominal response models for testlets, and a bivariate normal LV density. Due to the strong parametric assumptions made therein, we do not expect the baseline model to fit the data well. Inspired by the semiparametric fitting, we also specified an updated parametric model with nonlinear factor models with quintic mean functions for log-RT variables, 4PL models for item responses, nominal models for testlets, and a two-component normal mixture density for the LVs. Even though the updated model has yet to attain a fit comparable to the semiparametric model, it reproduces key functional patterns in the semiparametric estimates of the bivariate LV density and the conditional densities for the

MVs. Despite being tangential to the specific aims of the present work, these additional analyses exemplify another standard usage of semiparametric/nonparametric models: to provide diagnostic information about model-data fit and to guide model modification.

3.3. Detailed Configuration

For replicability, we provide all the tuning details involved in our analysis. PML estimation of the semiparametric simple structure model was implemented in the R package `spfa`, which can be downloaded at https://github.com/wwang1370/spfa and https://cran.r-project.org/web/packages/spfa/index.html.

Estimation $K = 13$ B-splines basis functions were used for approximating smooth functions defined on the unit interval. Each log-RT variable was linearly transformed to $[0, 1]$ using the sample minimum and maximum. The reference level for LVs and continuous MVs was set to $x_0 = y_0 = 0.5$; for discrete MVs, the reference level was set to the first response category $y_0 = 0$. We impose likelihood-ratio monotonicity on item CM442Q02 since both its responses and RT show the highest correlations with totals (see Tables 1 and 2). Intractable integrals appeared in the conditional densities (Eq. 1) were approximated by a 21-point Gauss-Legendre quadrature rescaled to the unit interval. The marginal likelihood function (Eq. 17) involves a two-dimensional integral over the unit square and was approximated by a tensor-product Gauss-Legendre quadrature. In each fitting, we executed the EM algorithm until the change in the penalized log-likelihood (i.e., Eq. 22) was less than 10^{-3} between consecutive iterations.

Penalty Weight Selection We selected the three penalty weights $\lambda_{(c)}$, $\lambda_{(d)}$, and $\lambda_{(g)}$ from the following sequences of decreasing values:

$$\begin{aligned}\lambda_{(c)} &\in \{10^{-1}, 10^{-2}, \ldots, 10^{-6}\},\\ \lambda_{(d)} &\in \{10^{1}, 10^{-1}, \ldots, 10^{-4}\},\\ \lambda_{(g)} &\in \{10^{-2}, 10^{-4}, \ldots, 10^{-8}\}.\end{aligned}$$

The empirical risk was computed by five-fold cross-validation (Eq. 23 with $S = 5$). The smoothest solution within one SE (estimated by Eq. 24) from the minimal-risk solution was deemed optimal.

Inference Conditional on the optimal penalty weights, we resampled $B = 100$ times with replacement, refit the model in each bootstrap sample, and examine the bootstrap distributions of fitted densities and model fit statistics. When computing fit diagnostics and summary statistics, we approximated intractable integrals by the same quadrature systems that were used in parameter estimation. The MV scoring function[5] for testlet responses was defined by $\varsigma_j(0) = 0$, $\varsigma_j(1) = \varsigma_j(2) = 1$, and $\varsigma_j(3) = 2$.

4. Results

4.1. Model Fit and Modification

In the initial fitting of the semiparametric simple-structure model (using all 26 MVs), our cross-validation procedure selects 10^{-4}, 10^{-1}, and 10^{-4} as the respective optimal values for $\lambda_{(c)}$, $\lambda_{(d)}$, and $\lambda_{(g)}$. A graphical display of the results can be found in the first row of Fig. 1.

Based on a full-data fitting with the optimal penalty weights, we summarize the residual correlation statistics (Eq. 25) for all pairs of MVs in a graphical table (Fig. 2). It is observed that dependencies within RT variables are well explained by the slowness factor, and similarly

[5]Note that this scoring function was also applied before computing the item-total correlation statistics in Table 2.

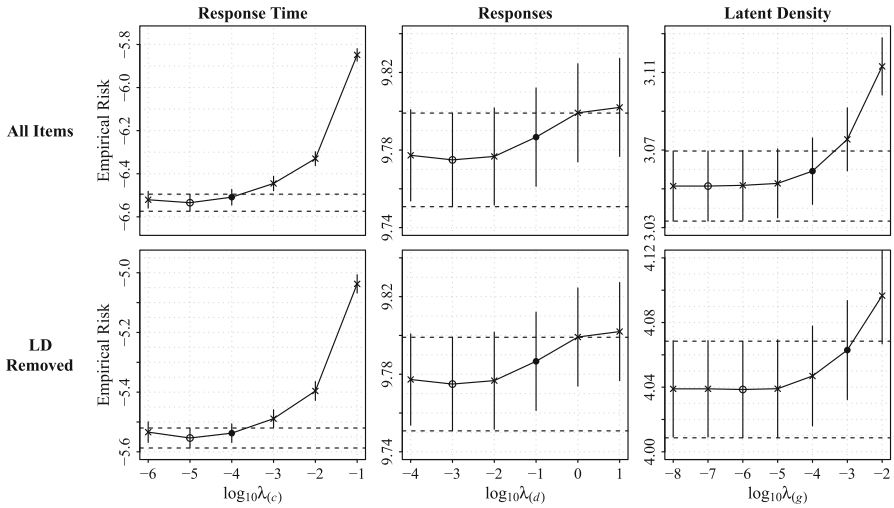

FIGURE 1.
Empirical risks (Eq. 23) and standard errors (SE; Eq. 24). The rows of the graphical table correspond to the initial fitting (with all items) and the updated fitting (without the response time of CM034Q01 and CM571Q01). The columns represent the three stages of penalty weight selection (see Sect. 2.3). Within each panel, empirical risk values are plotted as functions of based-10 log-transformed penalty weights. Vertical bars indicate one SE above and below the empirical risk. The minimized empirical risks are shown as circles, while the optimal solutions determined by the "one SE rule" were highlighted as filled dots. The band formed by two horizontal dashed lines indicates the one-SE region associated with the minimum empirical risk. Note that the two graphs in the second column are identical: This is because all item responses are retained, and thus we do not need to re-select $\lambda_{(d)}$. LD: Local dependence. $\lambda_{(c)}$, $\lambda_{(d)}$, $\lambda_{(g)}$: Penalty weights for continuous manifest variables (MVs), discrete MVs, and the latent density.

dependencies within item responses are well explained by the ability factor. The largest residual correlation in the left panel of Fig. 2 is 0.1 (between the log-RT of CM571Q01 and CM603Q01) with a 90% bootstrap CI [0.08, 0.12]. In contrast, we identify some non-ignorable residual dependencies between the log-RT and response of the same item (i.e., diagonal entries in the right panel of Fig. 2). The within-item residual correlations reach 0.14 (with a bootstrap CI [0.12, 0.15]) for both items CM034Q01 and CM571Q01. We also find a large negative residual correlation for item CM423Q1: The point estimate is -0.12, but the associated bootstrap CI $[-0.13, -0.1]$ covers -0.1. Meanwhile, the RT-response dependencies are well explained between items: The off-diagonal statistics in the right panel of Fig. 2 ranges between -0.07 and 0.08.

Given the above findings, we conclude that a simple factor structure largely suffices for modeling the item responses and RT in the 2015 PISA mathematics data. For two out of 13 items (CM034Q01 and CM571Q01), however, the associations between item-level response speed and accuracy are not fully addressed by individual differences in general processing speed and ability. To be clear of adverse impact caused by unaccounted residual dependencies, we dropped the log-RT variables for items CM034Q1 and CM571Q01 while letting their responses stay, which results in a modified simple-structure model with $m_1 = 11$ continuous MVs and $m_2 = 13$ discrete ones. Steps 1–3 (see Sect. 3.2) were repeated. The optimal $\lambda_{(c)}$ remains to be 10^{-4}, whereas the optimal $\lambda_{(g)}$ increases to 10^{-3} (see the second row of Fig. 1); the optimal $\lambda_{(d)} = 10^{-1}$ is retained as no change has been made to the item response variables. There is no more large residual this time. The ranges of the residual correlations are $[-0.04, 0.08]$ among log-RT variables, $[-0.03, 0.02]$ among response variables, and $[-0.11, 0.08]$ across responses and log-RT. Similar to the initial fitting, the only residual correlation beyond ± 0.1 is observed between the response and log-RT of item CM423Q1; however, the 90% bootstrap CI of the statistic is $[-0.13, -0.09]$ which contains -0.1. Therefore, we proceed to interpret the fitted densities based on the updated fitting.

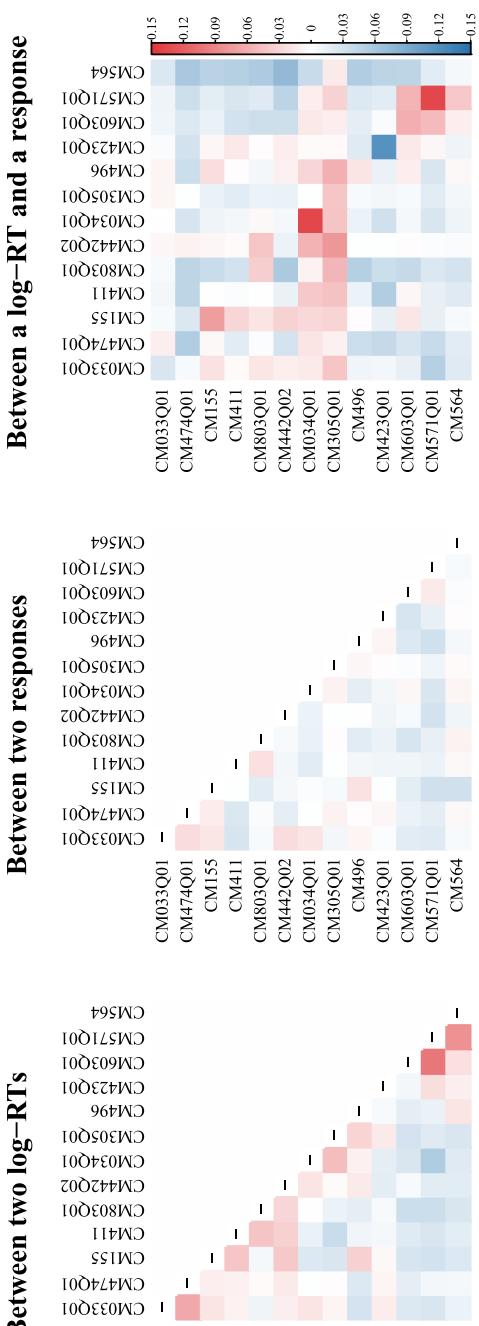

FIGURE 2.
Residual correlation statistics for the initial fitting of the semiparametric simple-structure model. Left: Residual correlations between two log-transformed response time (log-RT) variables. Middle: Residual correlations between two item/testlet responses. Right: Residuals between a log-RT variable and a item/test response, in which rows represent log-RT variables and columns represent responses. Positive residual correlations are shown in red, while negative residual correlations are shown in blue. A darker color indicates a larger magnitude (Color figure online).

4.2. *Conditional Densities of Manifest Variables*

Estimated conditional densities and means of the log-RT variables given the slowness factor are plotted in Fig. 3. Two major patterns are of interest here. First, although the high and low ends of the LV scale roughly map onto the longest and shortest RT for a majority of items/testlets, which justifies our decision to label the LV as "slowness", the conditional mean function appears to decrease at the high end for all items/testlets except for CM442Q02, on which we impose the monotonicity constraints (Eq. 8). However, we often cannot distinguish the observed downward trend from a flat one due to large sampling variability, which is manifested by wider bootstrap confidence bands in those areas. For item CM603Q01 and testlet CM564, the downturn at the high end cannot be explained away by sampling variability. It implies that, among slow responders for the first nine items/testlets, the slower they respond to the first nine the faster they tend to response to the last two. The second observation concerns the dips in conditional mean functions when the latent slowness is between -1 and 0. Taking sampling variability into account, the dips are not substantial for CM603Q01 and CM564; also recall that the conditional mean function was forced to be non-decreasing for item CM442Q02. As such, the observed dips reflect a negative association between the above triplet and the remaining items/testlets for the subset of respondents whose latent slowness values fall slightly below average.

Per a referee's request, we also examine the relationship between item-level RT and the ability factor. In our simple structure model, the log-RT variables Y_{ij}, $j = 1, \ldots, m_1$, do not directly load on the ability factor X^*_{i2}. Nevertheless, it remains possible to characterize the predictive distribution $Y_{ij}|X^*_{i2}$ by combining the conditional distribution of the slowness factor given the ability factor, i.e., $X^*_{i1}|X^*_{i2}$, with the conditional distribution $Y_{ij}|X^*_{i1}$ (shown in Fig. 3). Such RT-ability associations turn out to be weak in the present data set; detailed results can be found in the supplementary document.

Estimated item/testlet response functions are displayed in Fig. 4. Due to the large penalty weight (i.e., 10^{-1}), the fitted curves are smooth. For dichotomous items, the estimated curves for category 1 (i.e., correct answer) are largely in S-shape and typically have a restricted range (narrow

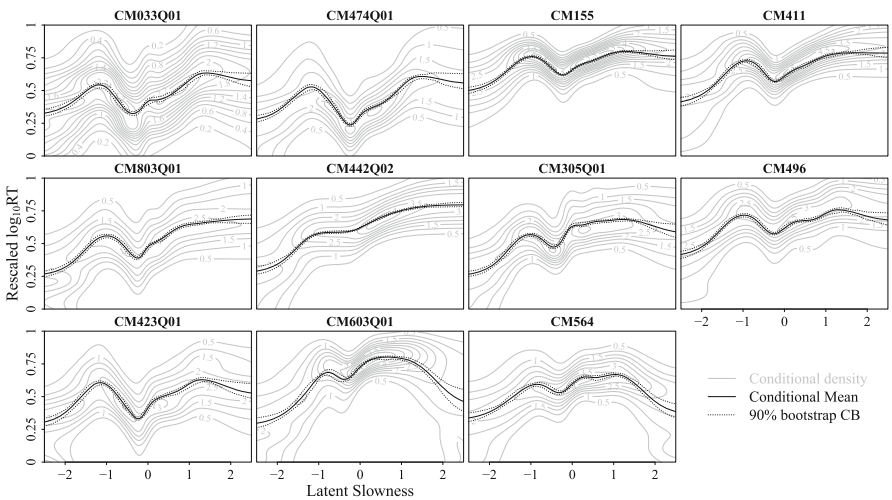

FIGURE 3.
Estimated conditional densities and means for log-10 response time (RT) variables (rescaled to [0, 1]). Each panel corresponds to a single item/testlet. Conditional densities of manifest variables given the slowness factor are visualized as contours in gray. Estimated conditional means are superimposed as solid curves in black. Dotted lines represent 90% bootstrap confidence bands (CBs) for estimated conditional mean curves.

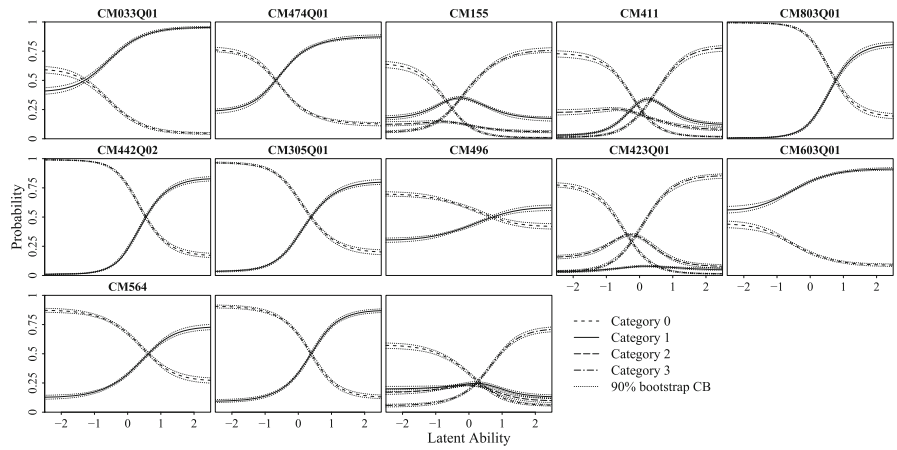

FIGURE 4.
Estimated conditional densities for discrete response variables, also known as item response functions (IRFs). Each panel corresponds to a single item/testlet. Curves for different categories are shown in different line types. Dotted lines represent 90% bootstrap confidence bands (CBs) for estimated IRFs.

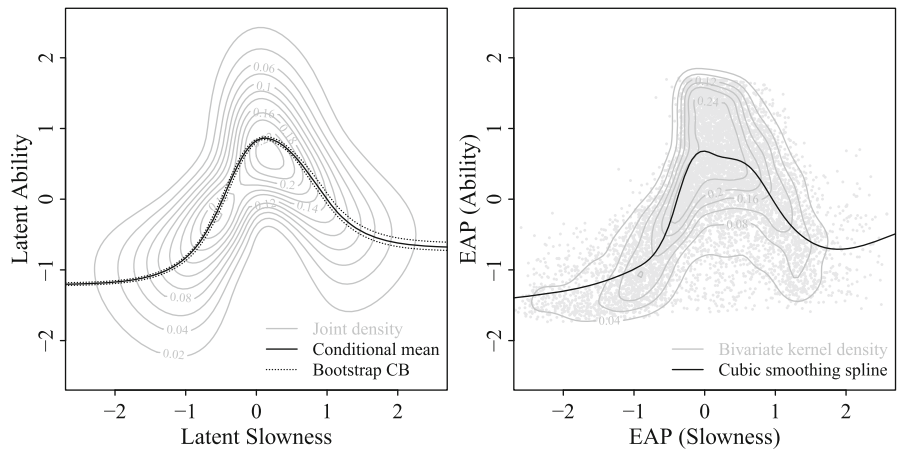

FIGURE 5.
Left: Estimated joint density for the slowness and ability factors (contours in gray) and the conditional mean of ability given slowness (black solid curve). Dotted lines represent 90% bootstrap confidence bands (CBs) for the estimated conditional mean curve. Right: Scatter plot for the expected *a posteriori* (EAP) scores of ability and slowness. A bivariate kernel density estimate (gray solid contours) and a smoothing spline regression line (black solid curve) are superimposed.

than the entire interval [0, 1]). Similarly, estimated testlet response functions for the first and last categories also appear to have (often different) upper asymptotes. Some items, e.g., CM305Q01 and CM423Q01, are poorly discriminating, manifested by relatively flat IRFs.

4.3. Latent Density and Scores

A contour plot for the estimated two-dimensional LV density, which is computed from the estimated B-spline copula density with standard normal marginals (Eq. 13), is provided in the left panel of Fig. 5. It is observed that high ability respondents tend to response in a moderate speed, whereas low ability respondents can respond either very rapidly or very slowly. The shape of the density contours is nowhere near elliptical, which calls the standard practice of fitting a bivariate

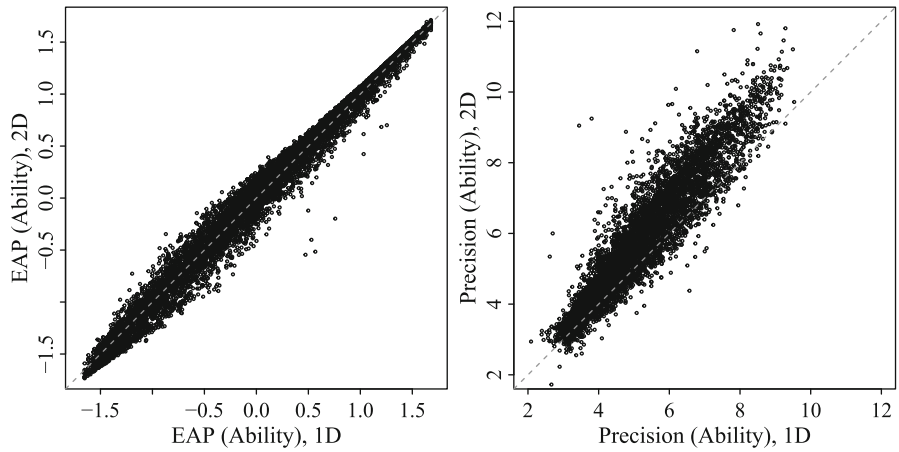

FIGURE 6.
Comparing expected a posteriori (EAP) scores for ability (left) and the associated predictive precisions (right) between the one-dimensional (1D) response-only model and the two-dimensional (2D) simple-structural model. In both panels, the dashed diagonal line in gray indicates equality.

normal LV density into question. A better parameterization of the latent density for this data would be a mixture of two bivariate normals—one with a positive correlation for fast responders (i.e., slowness < 0) and the other with negative correlation for slow responders (i.e., slowness > 0). A similar pattern is observed when we plot the ability EAP scores against the slowness EAP scores (right panel of Fig. 5), with an exception that EAP scores tend to be less variable than the true LVs.

To better visualize the relationship between the two latent factors in the population, we also plot the conditional mean of ability given slowness (i.e., the black solid curve in the left panel of Fig. 5)—in other words, a nonlinear regression that predicts ability by slowness. The η^2 statistic (Eq. 31) of the population nonlinear regression is 0.45 with a 90% bootstrap CI [0.44, 0.47], indicating a strong association (Cohen, 1988, Chapter 9). Stated differently, knowing respondents' processing speed on average reduces the uncertainty (measured by variance) in their mathematics ability by 45%. Recall that Zhan et al. (2018) reported a correlation of -0.2 between the speed (i.e., the reversal of slowness) and ability factors assuming bivariate normality, which implies $\eta^2 = (-0.2)^2 = 0.04$. The divergent conclusion reached by Zhan et al. (2018) is likely attributed to the restrictive parameterization of their measurement model: They forced the LV density to be bivariate normal and thus failed to capture the nonlinear relationship. In addition, a smoothing spline regression fitted to the EAP scores (i.e., the black sold curve in the right panel of Fig. 5) suggests a similar predictive relationship: The observed multiple R^2 statistic is 0.54, even higher than the population η^2.

As slowness/speed is a useful predictor of ability, it is anticipated that incorporating item-level RT information may improve the precision of IRT scale scores. Inspired by Bolsinova and Tijmstra (2018), we compare ability scores from the two-dimensional simple-structure model to those from the unidimensional semiparametric IRT model fitted to only responses in terms of their predictive precision (Sect. 2.5). It is first noted that the two sets of EAP scores are almost perfectly correlated (sample Pearson's correlation > 0.99; see the left panel of Fig. 6). We then plot the predictive precisions associated with the two sets of EAP scores in the right panel of Fig. 5. Because the test is short and some items (e.g., items CM305Q01 and CM423Q01) have low discriminative power (manifested by flat item response functions), the predictive precisions are not high in general. Pooling across the entire sample, the mean predictive precision based on the unidimensional model is 4.68 with an interquartile range (IQR) [3.44, 5.71], and the median

TABLE 3.
Predictive precisions of ability scores in quintile groups.

	Quintile groups (slowness)					Quintile groups (ability)				
	1	2	3	4	5	1	2	3	4	5
Avg prec (1D)	4.02	5.01	4.70	4.84	4.82	3.28	4.78	6.55	5.22	3.55
Avg prec (2D)	4.27	5.28	5.03	5.47	5.72	3.46	5.43	7.60	5.63	3.64
Improvement (in %)	5.99	5.31	7.18	12.98	18.64	5.48	13.47	15.89	7.96	2.52

Groups are determined by the slowness (left columns) and ability (right columns) scores computed from the two-dimensional simple-structure model.
Avg Prec: Average predictive precision within each group. 1D: One-dimensional model. 2D: Two-dimensional model.

predictive precision based on the two-dimensional simple-structure model is 5.15 with an IQR [3.57, 6.45]. That is to say, using the two-dimensional model improves the predictive precision for ability scores by 10.1% on average.

To assess scoring precision at different slowness and ability levels, we split the sample into quintile groups by the slowness and ability EAP scores (from the two-dimensional model), respectively. A group-by-group summary of scoring precisions is provided in Table 3. When groups are formed by slowness scores, more increases in precision are typically observed in higher quintile groups; the percentage of improvement can be as high as 18.64% in the fifth quintile group. In contrast, the largest improvement is attained in the middle quintile group (15.89%) when groups are determined by ability scores; the one-dimensional ability scores in the fifth quintile group are almost as precise as the two-dimensional scores.

5. Discussion

In the present paper, we perform a joint factor analysis for item response and RT data from the 2015 PISA mathematics assessment. In line with many previous studies that handled this type of data, our model features a simple factor structure with two LVs: The ability factor is indicated solely by item responses, the slowness factor is indicated solely by log-transformed RT variables, and the two LVs are permitted to covary in the population of respondents. The unique contribution of our work lies in the use of a semiparametric measurement model: We do not impose any restrictive functional forms of dependencies or distributional assumptions above and beyond the simple factor structure. Our model therefore fits the best to the data insofar as a simple factor structure is deemed proper. We approximate the functional parameters in the semiparametric factor model by cubic splines and estimate the resulting coefficients by PML: The penalty weights involved in the objective function are empirically selected via cross-validation. Inferences about model fit statistics and estimated functional parameters are conducted based on (nonparametric) bootstrap.

5.1. Implications

The semiparametric fitting reveals novel patterns that have yet been noticed in the existing literature, which has profound implications on the use of RT information in large-scale educational assessment.

First, a simple factor structure for ability and slowness fits reasonably well to the 2015 PISA mathematics data. Only two pairs of MVs exhibit excessive dependencies that are not well explained by the simple-structure model: Both pairs comprise the response and RT of the same

item. Furthermore, including or excluding the RT variables of the two flagged pairs is inconsequential for model-based inferences. Our finding verifies the prevalent psychometric theory that between-person heterogeneity in item response behaviors are reflections of individual differences in ability and general processing speed. However, the existence of within-item local dependence between responses and RT, albeit not influential for the current analysis of the PISA data, should be reassessed in other applications of simple-structure factor models.

Second, commonly used parametric factor models are too simple to fully capture the MV-LV relations. Our semiparametric model implies that the conditional means of log-transformed RT variables are generally increasing but nonlinear functions of the slowness factor; the conditional variances appear to be non-constant for some items too. The most commonly used log-normal RT model, however, implies a linear conditional mean and a constant conditional variance and thus is evidently misspecified. As Liu and Wang (2022) also reported in that the log-normal RT model fits substantially worse than the semiparametric model in a different empirical example, cautions are advised in choose a suitable measurement model for item-level RT. Meanwhile, a large penalty weight is selected for the semiparametric IRT model, and consequently the fitted IRFs are smooth. While the shapes of the IRFs closely resemble logistic curves, the presence of lower and upper asymptotes hints at a 4PL model (Barton & Lord, 1981) , rather than the more popular 1PL and 2PL models in psychometric operations.

Third, the ability and slowness factors are strongly associated, which is probably the most surprising observation since a weak correlation was reported in Zhan et al.'s (2018) analysis of the same data. The disparate finding of ours is ascribed to the use of a nonparametric latent density estimator, whereas the LV density is by default assumed to be (multivariate) normal in the vast majority of factor analysis applications. It then merely echoes a well-known fact that overly restrictive assumptions may lead to poorly fitting models and subsequently biased inferences. Diagnostics for non-normal LVs and measurement models equipped with non-parametric LV densities should be added to the routine toolbox for psychometricians. Future research is encouraged to examine the extent to which nonlinear factor models with non-normal latent densities can be beneficial in other assessment contexts.

Fourth, including item-level RT in the measurement model improves the precision of ability scores, which is an expected consequence as the ability factor can be well predicted by the slowness factor. While RT carries additional information about respondents' ability, induced by the association between ability and general processing speed, it remains unclear whether RT should be officially used for scoring purposes in high-stake educational assessment. On the one hand, the joint factor model estimated in the present paper results in about 10% increase in predictive precisions for ability scores on average. Adaptive tests based on such a joint factor model may need fewer test items to reach the desired measurement precision, leading to more cost-effective test administrations. On the other hand, the same measurement model may no longer hold once the respondents are aware that response speed somehow affects their performance scores. In the latter case, a re-calibration of the joint factor model and a re-evaluation on the usefulness of RT information are necessary.

5.2. Limitations

There are also a number limitations to be addressed by future investigation.

First, the selection of penalty weights by multifold cross-validation is time consuming. A referee suggested that computing a one-sample estimate of cross-validation error (e.g., Akaike information criterion; AIC) or a large-sample approximation to the Bayesian marginal log-likelihood (e.g., Bayesian information criterion; BIC) is computationally advantageous. For nonparametric/semiparametric models using penalized smoothing splines, however, we must substitute a properly defined "effective degrees of freedom (edf)" for the number of parameters in the usual

formulas of those information criteria. The *ad hoc* definition of edf proposed by Liu et al. (2016) for semiparametric IRT modeling can potentially be extended to the present context; however, the performance of the resulting information criteria in penalty weight selection remains unclear and should be investigated in future work.

Second, the sequential selection of multiple penalty weights does not guarantee that a globally optimal combination is found—it was only implemented as a workaround to alleviate the computational burden. Meanwhile, simultaneous selection on an outer-product grid (cf. Liu et al., 2016) suffers from the "curse of dimensionality" and may be computationally inviable when the total number of penalty weights to be selected is large. Future research is encouraged to apply and evaluate optimization-based penalty weight selection, such as the "performance-oriented iteration" by Gu (1992), to semiparametric factor analysis. With the aid of optimization-based selection, it is also possible to explore the feasibility of selecting different penalty weights for different MVs, which further enhances the flexibility of the model.

Third, some of our decisions regarding locally dependent MVs can be refined. While coding each testlet response pattern as a unique category does not lead to any information loss, treating the summed RT within a testlet as a single MV does. In addition, we remove within-item local dependencies between responses and RT by simply excluding the RT variables. Although our treatments suffice for the purpose of the current analysis, it is natural to seek extensions of the proposed model to handle local dependencies in a more elegant way. In our opinion, the best strategy to approach a pair of locally dependent MVs is to directly model their bivariate conditional distribution given the LVs. For example, we may express the joint density of two log-RT variables, say $Y_{ij} = y$ and $Y_{ij'} = z$, given the latent slowness variable $X_{i1} = x$ using a logistic density transform with a three-way fANOVA decomposition (Gu, 1995, 2013) :

$$f(y, z|x) \propto \exp\left(g^{y}(y) + g^{z}(z) + g^{xy}(x, y) + g^{xz}(x, z) + g^{yz}(y, z) + g^{xyz}(x, y, z)\right). \quad (34)$$

Equation 34 involves six functional components, each of which can be approximated via basis expansion under suitable side conditions. Despite the straightforward formulation, simultaneous estimation of a large number of functional parameters proves to be computationally challenging.

Fourth, a referee made an important point that the residual correlation statistic (Eq. 25) only captures linear dependencies, which does not rule out the existence of nonlinear residual dependencies and is a major limitation of our diagnostic procedure. There exist various measures for nonlinear associations: Recent example include the Hellinger correlation (Geenens & Lafaye de Micheaux, 2022) and the Wasserstein dependence coefficient (Mordant & Segers, 2022; see also Chatterjee, 2022, for a review). However, those measures are often less intuitive to interpret as no common rules of thumb have been developed. As an alternative, one may fit an extended semiparametric factor model with bivariate conditional densities (Eq. 34) and identify nonlinear dependencies from graphical displays of estimated conditional densities.

Fifth, the proposed semiparametric factor model can be generalized in a number of ways. Sometimes, multiple latent constructs are simultaneously measured by an instrument (e.g., personality assessment); hence, a joint factor analysis of responses and RT for those measures involves at least three LVs. Such extensions of the current semiparametric simple-structure model suffers from a two-fold "curse of dimensionality": The number of tensor-product basis functions grows exponentially when the dimension of a functional parameter's domain increases, and the number of tensor-product quadrature points for likelihood approximation also increases exponentially as the dimension of LVs increases. While the EM algorithm with numerical quadrature can be replaced by stochastic approximation (Cai, 2010a,b; Gu & Kong, 1998) to handle models with higher-dimensional LVs, reduced fANOVA parameterizations for conditional densities (Gu, 1995, 2013) and hierarchical formulations of B-spline copula (Kauermann et al., 2013) are handy for constructing economical approximations of multivariate functional parameters.

Sixth, resampling based procedures (e.g., bootstrap) are time consuming even if parallel processing via OpenMP (Dagum & Menon, 1998) is enabled in the current implementation of PML estimation. For parametric models, inferential procedures based on large-sample approximations fares more computationally efficient. However, it is generally more difficult to prove large-sample results for semiparametric/nonparametric models as the functional parameters are infinite dimensional. Theoretical foundations on the asymptotic theory for semiparametric/nonparametric measurement models have yet been established and are left for future research.

Last but not least, we emphasize that semiparametric approaches are better suited for analyses that are exploratory and data-driven in nature. There are also scenarios in which confirmatory and theory-driven model building is preferred: For instance, when the test is designed based on cognitive theory and administered in a controlled laboratory setting (e.g., the well known "mental rotation" example in the RT literature; Borst et al., 2011). One prominent example of theory-driven psychometrics is the integration of diffusion decision models with factor analysis (e.g., Kang et al., 2022, 2023a,b). Data-driven semiparametric models and theory-driven parametric models are both important yet mutually distinct tools to advance psychometricians' understanding in the role of processing speed in test-taking behavior.

Funding The work is sponsored by the National Science Foundation under grant No. 1826535.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability The dataset analyzed during the current study is available in the OECD PISA Database (https://www.oecd.org/pisa/data/).

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

References

Abrahamowicz, M., & Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika, 57*(1), 5–27.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series, 1981*(1), 1–8.

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling, 12*(4), 513–535.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459.

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika, 82*(4), 1126–1148.

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology, 69*(1), 62–79.

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology, 9*, 1525.

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics, 41*(2), 123–145.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology, 71*(1), 13–38.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology, 70*(2), 257–279.
Borst, G., Kievit, R. A., Thompson, W. L., & Kosslyn, S. M. (2011). Mental rotation is not easily cognitively penetrable. *Journal of Cognitive Psychology, 23*(1), 60–75.
Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*(1), 33–57.
Cai, L. (2010). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307–335.
Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.
Chatterjee, S. (2022). A survey of some recent developments in measures of association. arXiv preprint arXiv:2211.04702 .
Chen, Y., & Yang, Y. (2021). The one standard error rule for model selection: Does it work? *Stats, 4*(4), 868–892.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.
Currie, I. D., Durban, M., & Eilers, P. H. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(2), 259–280.
Dagum, L., & Menon, R. (1998). OpenMP: An industry standard API for shared-memory programming. *IEEE Computational Science and Engineering, 5*(1), 46–55.
De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102.
De Boor, C. (1978). *A practical guide to splines*. Berlin: Springer.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*(1), 1–22.
Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement, 58*(2), 281–303.
Dou, X., Kuriki, S., Lin, G. D., & Richards, D. (2021). Dependence properties of b-spline copulas. *Sankhya A, 83*(1), 283–311.
Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap*. Taylor & Francis.
Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 89–102.
Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika, 81*(2), 434–460.
Falk, C. F., & Cai, L. (2016). Semiparametric item response functions in the context of guessing. *Journal of Educational Measurement, 53*(2), 229–247.
Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series, 2015*(2), 1–17.
Geenens, G., & Lafaye de Micheaux, P. (2022). The hellinger correlation. *Journal of the American Statistical Association, 117*(538), 639–653.
Glas, C. A., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology, 63*(3), 603–626.
Goldhammer, F. (2015). Measuring ability, speed, or both? challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives, 13*(3–4), 133–164.
Gu, C. (1992). Cross-validating non-Gaussian data. *Journal of Computational and Graphical Statistics, 1*(2), 169–179.
Gu, C. (1995). Smoothing spline density estimation: Conditional distribution. *Statistica Sinica*, 709–726.
Gu, C. (2013). Smoothing spline ANOVA models. Springer.
Gu, M. G., & Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences, 95*(13), 7270–7274.
Gulliksen, H. (1950). *Theory of mental tests*. London: Wiley.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Berlin: Springer.
Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202.
Kang, H.-A. (2017). Penalized partial likelihood inference of proportional hazards latent trait models. *British Journal of Mathematical and Statistical Psychology, 70*(2), 187–208.
Kang, I., De Boeck, P., & Ratcliff, R. (2022). Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model. *Psychometrika*, 1–24.
Kang, I., Jeon, M., & Partchev, I. (2023). A latent space diffusion item response theory model to explore conditional dependence between responses and response times. *Psychometrika*, 1–35.
Kang, I., Molenaar, D., & Ratcliff, R. (2023). A modeling framework to examine psychological processes underlying ordinal responses and response times of psychometric data. *Psychometrika*, 1–35.
Kauermann, G., Schellhase, C., & Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical b-splines. *Scandinavian Journal of Statistics, 40*(4), 685–705.
Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence, 4*(4), 14.
Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359.
Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education, 2*(1), 1–24.

Liu, Y., Magnus, B. E., & Thissen, D. (2016). Modeling and testing differential item functioning in unidimensional binary item response models with a single continuous covariate: A functional data analysis approach. *Psychometrika, 81*, 371–398.
Liu, Y., & Wang, W. (2022). Semiparametric factor analysis for item-level response time data. *Psychometrika, 87*(2), 666–692.
Liu, Y., & Yang, J. S. (2018a). Bootstrap-calibrated interval estimates for latent variable scores in item response theory. Psychometrika, *83*(2), 333–354.
Liu, Y., & Yang, J. S. (2018). Interval estimation of latent variable scores in item response theory. *Journal of Educational and Behavioral Statistics*, *43*(3), 259–285.
Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.
McDonald, R. P. (1982). Linear versus models in item response theory. *Applied Psychological Measurement, 6*(4), 379–396.
Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement, 52*(1), 1–27.
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50*(1), 56–74.
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology, 68*(2), 197–219.
Mordant, G., & Segers, J. (2022). Measuring dependence between random vectors via optimal transport. *Journal of Multivariate Analysis, 189*, 104912.
Nelsen, R. B. (2006). *An introduction to copulas*. Berlin: Springer.
Nocedal, J., & Wright, S. (2006). *Numerical optimization*. New York: Springer.
OECD. (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris: PISA, OECD Publishing.
Pek, J., Sterba, S. K., Kok, B. E., & Bauer, D. J. (2009). Estimating and visualizing nonlinear relations among latent variables: A semiparametric approach. *Multivariate Behavioral Research, 44*(4), 407–436.
Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice, 35*(1), 38–47.
Ramsay, J. O., & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika, 56*(3), 365–379.
Ranger, J., & Kuhn, J.-T. (2012). A flexible latent trait model for response times in tests. *Psychometrika, 77*, 31–47.
Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling, 54*(2), 128.
Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics, 27*(3), 291–317.
Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement, 44*(5), 376–392.
Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology, 73*(3), 397–419.
Sklar, M. (1959). Fonctions de répartition àn dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris, 8*, 229–231.
Thissen, D., & Wainer, H. (2001). *Test scoring*. Taylor & Francis.
Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. Teachers College Bureau of Publications.
Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika, 2*(4), 249–254.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308.
van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*(1), 120–139.
van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*(5), 327–347.
van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 195–210.
von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6), 671–705.
Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*(1), 144–168.
Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics, 38*(4), 381–417.
Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement, 33*(2), 102–117.
Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement, 72*(2), 264–290.
Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology, 9*, 607.
Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics, 57*(3), 795–802.
Zhang, X., Wang, C., Weiss, D. J., & Tao, J. (2021). Bayesian inference for IRT models with non-normal latent trait distributions. *Multivariate Behavioral Research, 56*(5), 703–723.

Manuscript Received: 19 FEB 2023
Published Online Date: 16 NOV 2023