 doi:10.20944/preprints202105.0699.v1

Article

About the equivalence of the latent D-scoring model and the two-parameter logistic item response model

Alexander Robitzsch [1,2*]

[1] IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[2] Centre for International Student Assessment (ZIB), Kiel, Germany
* Correspondence: robitzsch@leibniz-ipn.de

Abstract: This article shows that the recently proposed latent D-scoring model of Dimitrov is statistically equivalent to the two-parameter logistic item response model. An analytical derivation and a numerical illustration are employed for demonstrating this finding. Hence, estimation techniques for the two-parameter logistic model can be used for estimating the latent D-scoring model. In an empirical example using PISA data, differences of country ranks are investigated when using different metrics for the latent trait. In the example, the choice of the latent trait metric matters for the ranking of countries. Finally, it is argued that an item response model with bounded latent trait values like the latent D-scoring model might have advantages for reporting results in terms of interpretation.

Keywords: latent D-scoring model, logistic item response model, identifiability, item parameter estimation, PISA

1. Introduction

Item response theory (IRT; [1]) is the statistical analysis of test items in education, psychology, and other fields of social sciences. Typically, a number of test items are administered to test takers and the interest is to infer the ability (performance or trait) of the test-takers. IRT models relate observed item responses to unobserved latent traits. Because the latent trait is unobserved, there are many plausible choices for modeling these relationships. The most popular class of IRT models are logistic IRT models [2]. Recently, in a series of papers, Dimitrov proposes an alternative IRT model, the so-called latent D-scoring model [3]. The main goal of this paper is to demonstrate that the newly proposed IRT model is statistically equivalent to the well-established two-parameter logistic IRT model.

The paper is structured as follows. In Section 2, IRT models are introduced in their general form. Afterward, the logistic IRT model and the latent D-scoring model are discussed. In Section 3, we show the statistical equivalence of the latent D-scoring model and the logistic IRT model utilizing an analytical derivation and a numerical illustration. Furthermore, we study the properties of the two models. Section 4 presents an empirical example that compares outcomes of the two different modeling strategies and compares them with two alternative parameterizations of the latent trait. Finally, the article closes with a discussion.

2. Item Response Modeling

In Section 2.1, we discuss the indeterminacy of the latent trait in IRT models. In Section 2.2, we focus on the logistic IRT model and its estimation. As an alternative IRT model, the latent D-scoring model is introduced in Section 2.3.

2.1. Indeterminacy of the Latent Trait in IRT Models

A unidimensional IRT model for dichotomous item responses $X_i \in \{0, 1\}$ is a statistical model [2]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

$$\mathrm{P}(\boldsymbol{X}=\boldsymbol{x})=\int_{-\infty}^{\infty}\prod_{i=1}^{I}\Big[P_i(\theta)^{x_i}(1-P_i(\theta))^{1-x_i}\Big]f(\theta)\,\mathrm{d}\theta \quad , \quad \theta\sim F \quad , \tag{1}$$

where f denotes the density function of the latent variable θ (also denoted as the latent trait) and $P_i(x,\theta)=\mathrm{P}(X_i=x|\theta)$ denotes the item response function of item i. Note that items $i=1,\ldots,I$ are conditionally independent given the latent trait θ. The model parameters in Equation (1) are typically not uniquely defined. Assume that one utilizes a monotone function $m:\mathbb{R}\to(0,1)$ for defining a transformed latent trait δ by $\delta=m(\theta)$. For example, m could be the logistic function $\Psi(x)=[1+\exp(-x)]^{-1}$ that maps the real line onto the unit interval $(0,1)$. Define $P_i^*(\delta)=P_i(m^{-1}(\delta))$, where m^{-1} denotes the inverse function of m. Furthermore, denote by g the density function of the transformed latent trait δ. The IRT model in Equation (1) can be equivalently written as

$$\mathrm{P}(\boldsymbol{X}=\boldsymbol{x})=\int_{0}^{1}\prod_{i=1}^{I}\Big[P_i^*(\delta)^{x_i}(1-P_i^*(\delta))^{1-x_i}\Big]g(\delta)\,\mathrm{d}\delta \quad . \tag{2}$$

The density g can be obtained from the density f by applying the density transformation theorem [4]

$$g(\delta)=\frac{f(m^{-1}(\delta))}{m'(m^{-1}(\delta))} \quad , \tag{3}$$

where $m'=\frac{\mathrm{d}m}{\mathrm{d}x}$ is the derivative of m with respect to θ.

It could be argued that only ordinal information can be extracted from the latent trait θ because the general IRT model (1) is only identified up to monotone transformations [5–8]. The indeterminacy of the latent trait metric implies that a researcher can seek a transformation $m(\theta)$ for the sake of enhancing interpretations of the results. One possible transformation is the true score metric $\tau=\tau(\theta)$ [2] that maps the θ metric from the real line to the bounded interval $(0,1)$ by defining

$$\tau(\theta)=\frac{1}{I}\sum_{i=1}^{I}P_i(\theta) \quad . \tag{4}$$

For a fixed value of θ, $\tau=\tau(\theta)$ is the expected value of the proportion of correctly solved items. Another alternative is the rank score metric $\rho=\rho(\theta)$ [8] that is defined by

$$\rho(\theta)=F(\theta) \quad , \tag{5}$$

where F is the distribution function of θ. One can show that ρ follows a uniform distribution (hence, the label "rank score"):

$$\mathrm{P}(\rho\leq u)=\mathrm{P}(F(\theta)\leq u)=\mathrm{P}(\theta\leq F^{-1}(u))=F(F^{-1}(u))=u \quad , \quad 0<u<1 \quad . \tag{6}$$

2.2. Logistic Item Response Model

An important class of IRT models is the class of logistic IRT models. Logistic IRT models employ the logistic link function for parameterizing IRFs. The IRFs in the two-parameter logistic (2PL) model [9] are given by

$$\mathrm{P}(X_i=1|\theta)=P_i(\theta)=\frac{1}{1+\exp(-a_i(\theta-b_i))}=\Psi(a_i(\theta-b_i)) \quad , \quad \theta\sim F \quad , \tag{7}$$

where a_i are item discriminations and b_i are item difficulties. The one-parameter logistic (1PL) model (Rasch model; [10]) is obtained by setting all item discriminations equal to one (i.e., $a_i=1$ for $i=1,\ldots,I$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

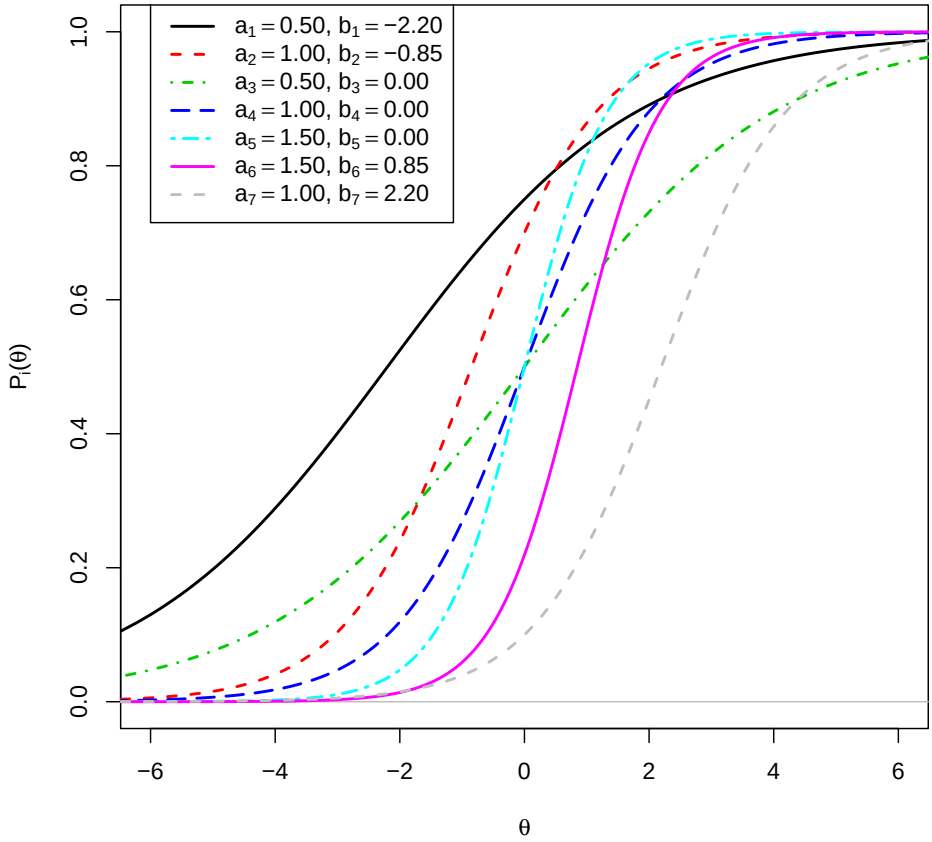

Figure 1. Item response functions for seven items of the 2PL model

In Figure 1, IRFs of seven items of the 2PL model are displayed (see the figure legend for item parameters a_i and b_i). It can be seen that items with higher item discriminations a_i have steeper slopes. Also, items with larger item difficulties b_i are shifted to the right. A fundamental property of IRFs in the 2PL model is a lower asymptote of zero and an upper asymptote of one. Hence, persons with very low abilities ($\theta \to -\infty$) have almost zero probability of correctly solving any item in the test, while highly able persons ($\theta \to \infty$) correctly solve items with a probability of one. Alternative IRT models allow lower and upper asymptotes different from 0 or 1, respectively [11].

In many applications, a normal distribution $\mathrm{N}(\mu, \sigma^2)$ for the latent trait θ is assumed [8]. However, more flexible distributions or semiparametric specifications are possible [12,13]. Identifications constraints are required in the 1PL and 2PL models for the estimation of model parameters. In the 1PL model, one can identify the model by setting $\mu = 0$ or fixing an item difficulty of a reference item to 0 (or to a prespecified value). Alternatively, one can constrain the sum of the item difficulties equal to zero. In the 2PL model, identification can be ensured by posing a standard normal distribution $\mathrm{N}(0, 1)$ (i.e., $\mu = 0$ and $\sigma = 1$). Alternatively, a reference item i_0 can be chosen for which $a_{i_0} = 1$ and $b_{i_0} = 0$ are used as fixed values in the estimation. Using a reference item has the advantage that the distribution F of θ can be flexibly estimated without using constraints on some parameters of F.

The 1PL model or the 2PL model can be estimated using marginal maximum likelihood (MML) or joint maximum likelihood (JML) estimation [2]. It is noteworthy that $\sum_{i=1}^{I} X_i$ is a sufficient statistic for θ in the 1PL model, while $\sum_{i=1}^{I} a_i X_i$ is the corresponding sufficient statistic in the 2PL model. Hence, the different models imply different interpretations implications of the trait because the contribution of items to the variable of interest differs considerably [14].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

2.3. Dimitrov's Latent D-Scoring Model

Dimitrov proposes an alternative IRT model that has a bounded metric for the latent trait. His latent D-scoring (LDS) model [3,15] includes a latent trait δ that takes values in the interval $(0,1)$. The IRF in the LDS model is given as [16]

$$P(X_i = 1|\delta) = \frac{1}{1 + \left[\frac{1-\delta}{\delta}\frac{\beta_i}{1-\beta_i}\right]^{\alpha_i}} \quad , \quad \delta \sim G \quad , \tag{8}$$

where G is some distribution on $(0,1)$. Item discriminations α_i are nonnegative, while item difficulties δ_i range between 0 and 1. The IRF in Equation (8) is also referred to as the rational function model with two item parameters [16]. The IRFs of the LDS model for seven items are shown in Figure 2 (see the figure legend for item parameters α_i and β_i).

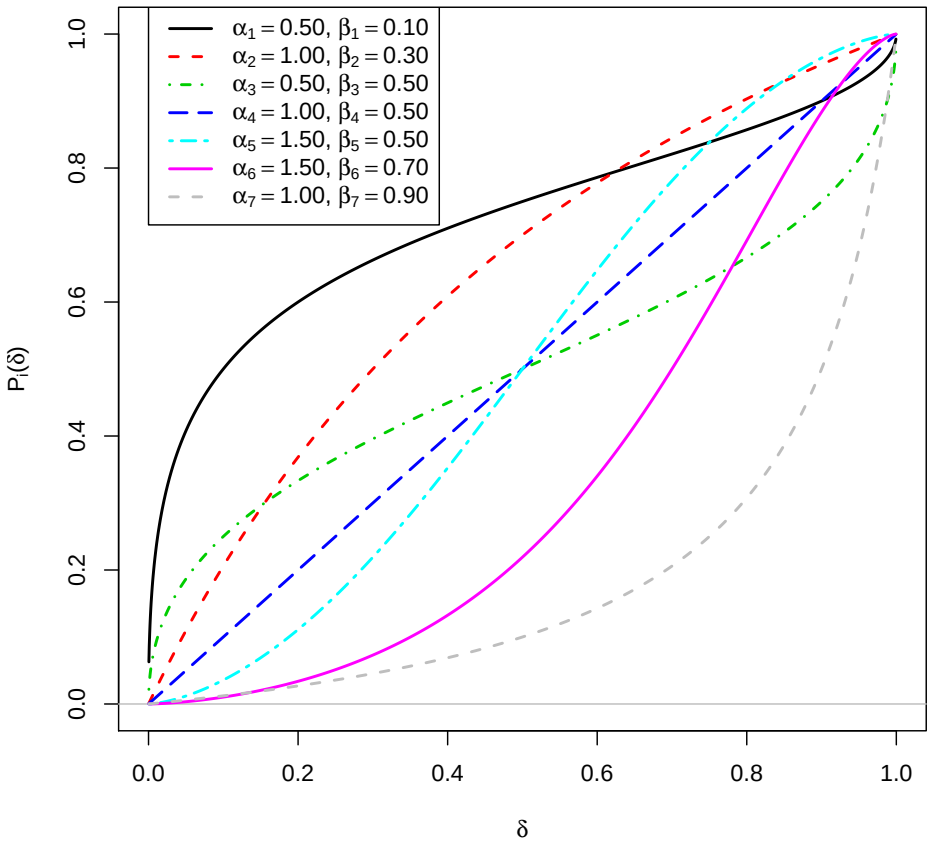

Figure 2. Item response functions for seven items of the LDS model

The LDS model with one item parameter is obtained by setting $\alpha_i = 1$:

$$P(X_i = 1|\delta) = \frac{1}{1 + \frac{1-\delta}{\delta}\frac{\beta_i}{1-\beta_i}} \quad , \quad \delta \sim G \quad . \tag{9}$$

In the following, we only consider the case of the LDS model with two item parameters and briefly discuss the special case of one item parameter in the discussion section.

The LDS model can be estimated with MML [3] or JML [17]. In Section 3, we show that identification constraints are needed for the estimation of the model. The latent D-scoring model is applied in psychometric areas of linking and equating [17], differential item functioning [15], and the development of multi-stage tests [18].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

3. Relation of the Latent D-Scoring Model and the 2PL Model

In this section, we show the close correspondence of the 2PL model and the LDS model. It is demonstrated that the two models are equivalent using analytical (Section 3.1) and numerical (Section 3.2) arguments. However, the two models imply different consequences regarding measurement precision and interpretations (Section 3.3). Finally, we propose an extension of the LDS model to multiple dimensions in Section 3.4.

3.1. Equivalence of the Latent D-Scoring Model and the 2PL Model

In this subsection, we analytically show that the LDS model is statistically equivalent to the 2PL model. Consequently, the model parameters of the 2PL model can be transformed to obtain model parameters of the LDS model.

The IRF in the 2PL model (Equation (8)) can be rewritten as

$$\mathrm{P}(X_i = 1|\delta) = \frac{1}{1 + \exp\left(-\alpha_i \left[\log \frac{\delta}{1-\delta} - \log \frac{\beta_i}{1-\beta_i}\right]\right)} \quad , \quad \delta \sim G \quad , \tag{10}$$

where G is the distribution function of δ. By defining $\theta = \log \frac{\delta}{1-\delta}$, $b_i = \log \frac{\beta_i}{1-\beta_i}$, and $a_i = \alpha_i$, one can rephrase the LDS model in Equation (10) as the 2PL model. Equivalently, we can write $\delta = \Psi(\theta) = [1 + \exp(-\theta)]^{-1}$ as the logistic transform of θ. Note that the logistic transform of $\delta = \Psi(\theta)$ was also discussed in [7,8]. Hence, the LDS model is just a reparametrized as the 2PL model. Hence, estimation routines for the 2PL model can be used for estimating the latent D-scoring model, and item parameters are transformed afterward; that is, $\alpha_i = a_i$ and $\beta_i = \Psi(b_i)$.

The distribution of δ can also be derived from the distribution of θ. The density function g of δ can be obtained from the density function f of θ by applying Equation (3)

$$g(\delta) = \frac{1}{\delta(1-\delta)} f\left(\log \frac{\delta}{1-\delta}\right) \quad . \tag{11}$$

Conversely, the density function θ can also be obtained from the density function of δ by

$$f(\theta) = \Psi(\theta)(1 - \Psi(\theta)) g(\Psi(\theta)) \quad . \tag{12}$$

The estimation of the LDS model using software for the 2PL model requires a correct specification of the distribution for θ. Suppose that a particular distributional assumption is posed on δ with density g. In that case, the estimation procedure must ensure that the assumed distribution for θ aligns with the implied density f for θ (see Equation (12)) to avoid biased item parameter estimates.

In Section 2.2, we showed that identification constraints are needed for estimating the 2PL model. Because the latent D-scoring model is equivalent to the 2PL model, the former also needs identification constraints. In the 2PL model, the location (i.e., the mean μ) and the scale (i.e., the standard deviation σ) for the latent trait θ can be fixed in the estimation. This would translate into identification constraints for the LDS model. Alternatively, a reference item i_0 could be chosen for the LDS model with fixed parameters $\alpha_{i_0} = 1$ and $\beta_{i_0} = 0.5$.

3.2. Numerical Illustration

This subsection demonstrates that the LDS model can be estimated using software for the 2PL model. We used item parameters of the LDS model that were also used in Figure 2 (see also Table 1). The multivariate distribution of $I = 7$ items according to the LDS model can be written as

$$P(\boldsymbol{X} = \boldsymbol{x}) = \int_0^1 \prod_{i=1}^{I} \left[P_i(\delta; \alpha_i, \beta_i)^{x_i} (1 - P_i(\delta; \alpha_i, \beta_i))^{1-x_i} \right] g(\delta) \, \mathrm{d}\delta \quad , \tag{13}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

where $P_i(\delta;\alpha_i,\beta_i)$ is the IRF for the ith item of the LDS model, and $\boldsymbol{x} = (x_1,\ldots,x_I)$. Note that there are $2^I = 128$ different item response patterns. The corresponding marginal probabilities $P(\boldsymbol{X} = \boldsymbol{x})$ are computed using (13) and numerical integration with respect to θ. We do not employ a simulation study because we try to show the statistical equivalence of the two models in a population, not in a sample.

We considered two distributions for δ. First, δ followed a beta distribution Beta(4,2) [19]. Second, δ followed a logit-normal distribution LogitN$(0.6, 1.2^2)$, that is $\theta = \log \frac{\delta}{1-\delta}$ is normally distribution with a mean of 0.6 and a standard deviation of 1.2 [20].

The 2PL model was estimated in the R [21] package sirt [22] using a sample weights option that inputs the item response pattern probabilities $P(\boldsymbol{X} = \boldsymbol{x})$. To avoid a restrictive distributional assumption on θ, we used a fixed grid of 61 equidistant θ values ranging between -6 and 6 and employed log-linear smoothing of the trait distribution up to the first four moments [13,23]. The item parameters of the fourth item were fixed (i.e., $a_4 = 1$ and $b_4 = 0$ in the 2PL model which corresponds to $\alpha_4 = 1$ and $\beta_4 = 0.5$ in the LDS model). The 2PL model was estimated using MML estimation and an EM algorithm [24].

Table 1. Estimated item parameters for the numerical illustration assuming a logit-normal distribution and a beta distribution

Item	LDS		2PL		LogitN$(0.6,1.2^2)$		Beta(4,2)	
	α_i	β_i	a_i	b_i	$\hat{\alpha}_i$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\beta}_i$
1	0.50	0.10	0.50	−2.20	0.501	0.101	0.505	0.102
2	1.00	0.30	1.00	−0.85	1.002	0.301	0.988	0.296
3	0.50	0.50	0.50	0.00	0.501	0.500	0.513	0.502
4‡	1.00	0.50	1.00	0.00	*1.000*	*0.500*	*1.000*	*0.500*
5	1.50	0.50	1.50	0.00	1.503	0.500	1.472	0.499
6	1.50	0.70	1.50	0.85	1.504	0.700	1.531	0.701
7	1.00	0.90	1.00	2.20	1.002	0.900	1.071	0.892

Note. LDS = latent D-scoring model; 2PL = two-parameter logistic model; LogitN = logit-normal distribution; Beta = beta distribution; ‡ = Item 4 was used as a reference item in estimation by fixing $a_4 = 1$ and $b_4 = 0$ (i.e., $\alpha_4 = 1$ and $\beta_4 = 0.50$).

Results for this numerical illustration are presented in Table 1. It can be seen that estimated item parameters $\hat{\alpha}_i$ and $\hat{\beta}_i$ for the LDS model almost perfectly recover true values in the case of the logit-normal distribution. This finding can be expected because the log-linear smoothing approach includes the normal distribution as a particular instance (smoothing up to two moments). Slightly larger deviations were observed if the distribution for δ was a beta distribution. The logit transform of the beta distribution is not correctly represented by a log-linear smoothing approach up to four moments which explains slight biases in item parameter estimates. For example, $\hat{\beta}_7 = 0.892$ deviated from the true value $\beta_7 = 0.90$ and $\hat{\alpha}_5 = 1.472$ deviated from $\alpha_5 = 1.50$. However, these numerical differences are probably negligible in practical applications and confirm our analytical reasoning for the equivalence of the 2PL and the LDS model.

3.3. Conditional Standard Errors for the Latent Trait

In this subsection, we study the amount of information for the latent trait that can be extracted with the 2PL model and the LDS model by using the concept of item information. Let $\boldsymbol{x}_{pi} = (x_{p1},\ldots,x_{pI})$ denote the vector of item responses of person p. For IRFs P_i (depending on already estimated item parameters), the maximum likelihood estimate $\hat{\theta}_p$ for the latent trait of person p is given as [1]

$$\hat{\theta}_p = \arg\max_{\theta} \sum_{i=1}^{I} \{x_{pi} \log P_i(\theta) + (1 - x_{pi}) \log(1 - P_i(\theta))\} \quad . \tag{14}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

Hence, the standard error associated with the estimate $\hat{\theta}_p$ is related to the information function that is obtained as the negative value of the second derivative of the log-likelihood function evaluated at $\hat{\theta}_p$. The information that is provided by item i in (14) is then given as

$$-x_{pi}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log P_i(\theta) - (1-x_{pi})\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(1-P_i(\theta)) \quad . \tag{15}$$

This allows defining the (expected) item information I_i for item i [25]

$$I_i(\theta) = -\pi_i\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log P_i(\theta) - (1-\pi_i)\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(1-P_i(\theta)) \quad , \tag{16}$$

where $\pi_i = \mathrm{E}(X_i)$ is the expected value for item i. In the literature, the observed item information

$$OI_i(\theta) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log P_i(\theta) \tag{17}$$

is often defined as the item information function. However, this function can become negative for some IRT models and the LDS model in particular [25], which is why preferring (16) for ensuring positivity of the item information function. For the 2PL model, the expected and observed item information coincide and are given as

$$I_i(\theta) = a_i^2 P_i(\theta)(1-P_i(\theta)) \quad . \tag{18}$$

Equation (18) implies that the least information is available for extreme θ values (i.e., extremely negative or positive).

The test information $I(\theta)$ is defined as $I(\theta) = \sum_{i=1}^{I} I_i(\theta)$. It quantifies the information that is provided by the test at each latent trait value θ. The conditional standard error for the latent trait θ is given by $\mathrm{SE}(\theta) = 1/\sqrt{I(\theta)}$.

One can similarly define the item information function for δ for the LDS model (see also [16]):

$$I_i(\delta) = -\pi_i\frac{\mathrm{d}^2}{\mathrm{d}\delta^2}\log P_i(\delta) - (1-\pi_i)\frac{\mathrm{d}^2}{\mathrm{d}\delta^2}\log(1-P_i(\delta)) \quad . \tag{19}$$

Analogously, the test information function $I(\delta) = \sum_{i=1}^{I} I_i(\delta)$ can be defined for the latent trait δ.

Because the latent D-scoring model is equivalent to the 2PL model (see Section 3.1), $\delta = \Psi(\theta)$ is a monotonous transformation of θ and the test information function for θ can be converted into the test transformation for δ. More generally, let $\delta = m(\theta)$ be a monotone differentiable transformation, the test information function for δ can be computed from the test information function for θ (see [2]):

$$I(\delta) = I(m(\theta)) = \frac{1}{m'(\theta)}I(\theta) \quad , \tag{20}$$

where $m' = \frac{\mathrm{d}m}{\mathrm{d}\theta}$. Equation (20) can be rewritten for conditional standard errors as

$$\mathrm{SE}(\delta) = \mathrm{SE}(m(\theta)) = \sqrt{m'(\theta)}\,\mathrm{SE}(\theta) \quad . \tag{21}$$

Hence, the conditional standard error $\mathrm{SE}(\delta)$ for the LDS model is given as

$$\mathrm{SE}(\delta) = \mathrm{SE}(\Psi(\theta)) = \sqrt{\Psi(\theta)(1-\Psi(\theta))}\,\mathrm{SE}(\theta) \quad . \tag{22}$$

In Section 3.2, we demonstrated that the LDS model is equivalent to the 2PL model. For the item parameters of the seven items used in the demonstration (see Table 1), the conditional standard errors for θ and δ are shown in Figure 3. It can be seen that the 2PL model measures the latent trait θ less precisely for extremely large negative and extremely large positive values, that is for low- and high-achieving persons. In line with the results

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

of [3], the converse holds for the LDS model. Conditional standard errors are smallest for persons with δ values near 0 or 1. Hence, statements about measurement precision in different ranges of values for the latent trait strongly depend on the chosen metric (see also [26]). Interestingly, the transformed latent trait $\xi = m(\theta) = \int_{-\infty}^{\theta} \sqrt{I(u)}\, \mathrm{d}u$ (the so-called arc length metric; see [7]) has homogeneous standard errors among the latent trait

$$\mathrm{SE}(\xi) = \mathrm{SE}(m(\theta)) = 1 \quad . \tag{23}$$

These observations indicate that it is difficult to state for which subgroups of persons adaptive or multi-stage testing [27] provides measurement precision gains because such statements depend on the chosen metric.

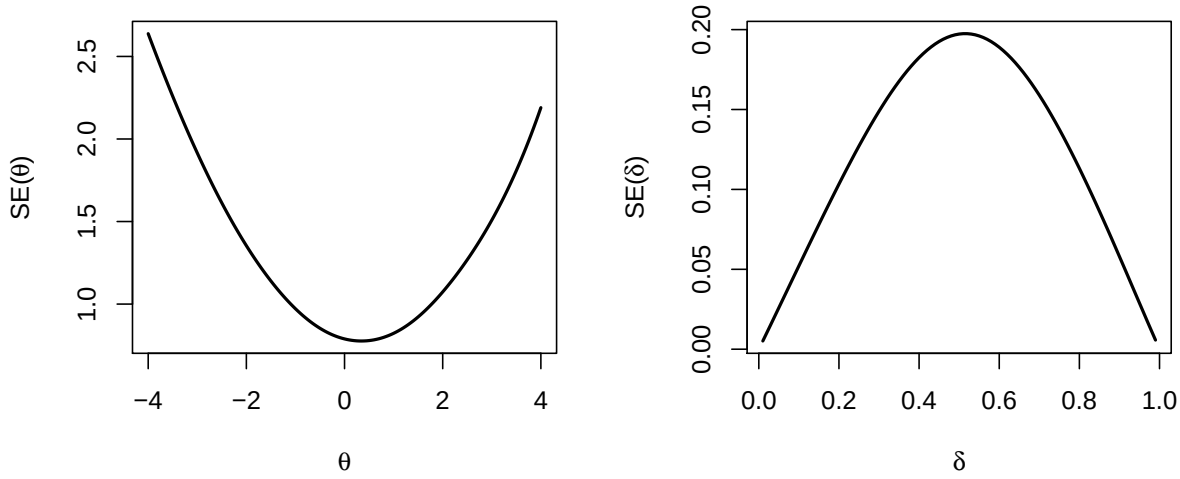

Figure 3. Conditional standard errors $\mathrm{SE}(\theta)$ for the 2PL model (left figure) and $\mathrm{SE}(\delta)$ for the LDS model (right figure)

3.4. A Multidimensional Latent D-Scoring Model

To our knowledge, the LDS model has only be investigated for a unidimensional latent variable δ. However, in applications, multidimensional traits are often of interest [28,29]. We now show that an apparent extension of the LDS model to multiple dimensions can be obtained by using the same transformations of the multidimensional variant of the 2PL model. We illustrate the arguments for two dimensions θ_1 and θ_2.

The multidimensional logistic IRT model can be written as [29]

$$\mathrm{P}(X_i = 1|\theta_1, \theta_2) = \frac{1}{1 + \exp(-a_{i1}\theta_1 - a_{i2}\theta_2 + d_i)} \quad , \quad (\theta_1, \theta_2) \sim F \quad , \tag{24}$$

where F is a bivariate distribution of (θ_1, θ_2) and θ_d $(d = 1, 2)$ attain values on the real line. Define transformed latent traits $\delta_d = \Psi(\theta_d) = [1 + \exp(-\theta_d)]^{-1}$ $(d = 1, 2)$ as the logistic transformations of θ_d. Like in the unidimensional LDS model, the δ_d variables attain values in the interval $(0, 1)$. Note that the inverse transformation is given as $\theta_d = \log \frac{\delta_d}{1-\delta_d}$. Then, employing the same strategy as in Section 3.1, one can rewrite Equation (24) by using $\beta_i = \Psi(d_i)$ and $\alpha_{id} = a_{id}$ as

$$\mathrm{P}(X_i = 1|\delta_1, \delta_2) = \frac{1}{1 + \left[\frac{1-\delta_1}{\delta_1}\right]^{\alpha_{i1}} \left[\frac{1-\delta_2}{\delta_2}\right]^{\alpha_{i2}} \frac{\beta_i}{1-\beta_i}} \quad , \quad (\delta_1, \delta_2) \sim G \quad . \tag{25}$$

Hence, the multidimensional 2PL model can easily be reparametrized for defining a multidimensional LDS model. The generalization to more than two dimensions is straightforward.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

4. Empirical Example: PISA 2006 Reading

4.1. Method

In order to illustrate the consequences of the choice of different metrics of the latent trait in multiple-group comparisons, we analyzed the data from the programme for international student assessment (PISA) conducted in 2006 (PISA 2006; [30]). In this situation, groups constitute countries. We included 26 countries (see Table 2) that participated in 2006 and focused on the reading test (see [31] and [32] for other studies using this dataset).

Items for the reading domain were only administered to a subset of the participating students. We included only those students who received a test booklet with at least one reading item. This resulted in a total sample size of 110,236 students (ranging from 2,010 to 12,142 students between countries). In total, 28 reading items nested within eight reading texts were used in PISA 2006. Six of the 28 items were polytomous and were dichotomously recoded, with only the highest category being recoded as correct.

In all analyses, student weights were taken into account. Within a country, student weights were normalized to a sum of 5,000, so that all countries contributed equally to the analyses.

In a first step, the 2PL model was estimated based on the data comprising students of all 26 countries. Student weights were taken into account, and a normal distribution was posed for θ in the estimation. The obtained item parameters $\hat{a}_i$ and $\hat{b}_i$ were fixed in the second step when estimating the trait distribution in each country. More concretely, the 2PL model was fitted using the R [21] package sirt [22] using MML estimation. The 2PL model was estimated by using a discrete grid of $T = 121$ equidistant θ points ranging between -6 and 6 for numerical integration of the involved integrals in the log-likelihood function of the 2PL model. As in Section 3.2, log-linear smoothing up to four moments of the trait distribution [13] within a country was employed to allow non-normal distributions. Assume that the estimated parametric distribution for θ in country g is $\pi_{gt} = P(\theta_t; \delta_g)$ for grid values θ_t $(t = 1, \ldots, T)$ and country-specific distribution parameters δ_g. Afterwards, individual posterior distributions $h_p(\theta_t | \boldsymbol{x}_p)$ $(t = 1, \ldots, T)$ were computed as

$$h_p(\theta_t|\boldsymbol{x}_p) = \frac{\prod_{i=1}^{I}\left[P_i(\theta_t;\hat{a}_i,\hat{b}_i)^{x_{pi}}\left(1-P_i(\theta_t;\hat{a}_i,\hat{b}_i)\right)^{1-x_{pi}}\right]\pi_{gt}}{\sum_{u=1}^{T}\prod_{i=1}^{I}\left[P_i(\theta_u;\hat{a}_i,\hat{b}_i)^{x_{pi}}\left(1-P_i(\theta_u;\hat{a}_i,\hat{b}_i)\right)^{1-x_{pi}}\right]\pi_{gu}} \quad , \tag{26}$$

where $P_i(\theta_t; \hat{a}_i, \hat{b}_i)$ is the IRF of item i from the 2PL model using estimated item parameters $\hat{a}_i$ and $\hat{b}_i$ from the total sample. By construction, it holds that $\sum_{t=1}^{T} h_p(\theta_t|\boldsymbol{x}_p) = 1$. For N_g persons per country g, the country means $\hat{\mu}_{\theta,g}$ on the logit metric θ was estimated by

$$\hat{\mu}_{\theta,g} = \frac{1}{W}\sum_{p=1}^{N_g} w_p \sum_{t=1}^{T} \theta_t h_p(\theta_t|\boldsymbol{x}_p) \quad , \tag{27}$$

where the person weights w_p sum to $W = 5,000$ within a country (i.e, $\sum_{p=1}^{N_g} w_p = W$). Country-specific standard deviations $\hat{\sigma}_{\theta,g}$ can be computed in a similarly:

$$\hat{\sigma}_{\theta,g} = \sqrt{\frac{1}{W}\sum_{p=1}^{N_g} w_p \sum_{t=1}^{T} \theta_t^2 h_p(\theta_t|\boldsymbol{x}_p) - \hat{\mu}_{\theta,g}^2} \quad . \tag{28}$$

Besides the logit metric θ, we also investigated the metric δ based on LDS model, the true score metric τ, and the rank score metric ρ. All three alternative metrics are monotone transformations $m(\theta)$ of θ. The country mean $\hat{\mu}_{m(\theta),g}$ at the transformed metric was calculated as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

$$\hat{\mu}_{m(\theta),g} = \frac{1}{W} \sum_{p=1}^{N_g} w_p \sum_{t=1}^{T} m(\theta_t) h_p(\theta_t | \boldsymbol{x}_p) \quad . \tag{29}$$

Using (29), the standard deviation of $m(\theta)$ can be computed in a similarly to (28). Furthermore, conditional standard errors for the four latent trait metrics are computed for the whole sample containing all students. The item information is obtained by using the second derivatives of IRFs with respect to the metrics θ, δ, τ, and ρ (see Equation (16)).

4.2. Results

In Table A1 of Appendix A, estimated item parameters $\hat{a}_i$ and $\hat{b}_i$ from the 2PL model are shown. These item parameters were transformed into parameters of the equivalent LDS model (see columns $\hat{\alpha}_i$ and $\hat{\beta}_i$ in Table A1). The IRFs of seven selected items are displayed in Figure A1 in Appendix A for the four latent trait metrics θ, δ, τ and δ. IRFs for the bounded metrics δ, τ and δ look very similar.

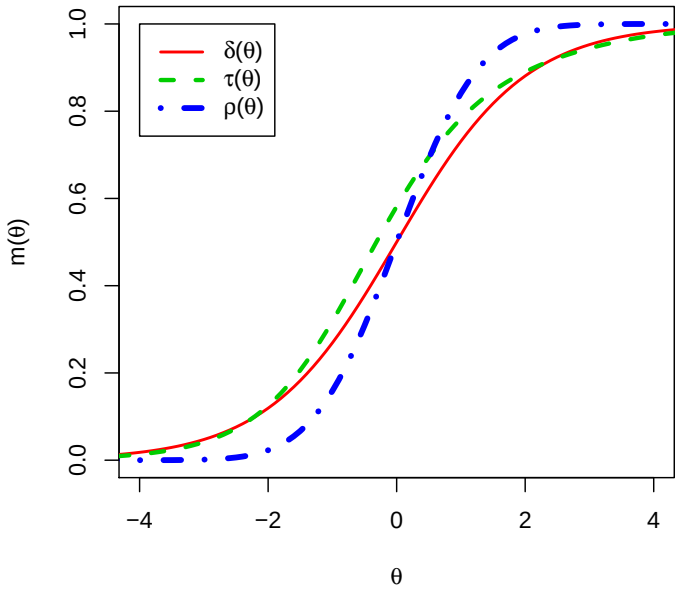

Figure 4. Transformation functions $\delta = \delta(\theta)$, $\tau = \tau(\theta)$ and $\rho = \rho(\theta)$ of latent ability θ for the PISA 2006 reading test

In Figure 4, the transformation functions $\delta = \delta(\theta)$, $\tau = \tau(\theta)$ and $\rho = \rho(\theta)$ are depicted. The latent D-score δ and the true score τ follow a very close transformation function. The rank score ρ differs from the former two in the tails of the θ distribution. Hence, it can be expected that δ and τ provide similar country rankings, while using ρ might lead to slightly different country rankings.

In Figure 5, conditional standard errors are displayed. It can be seen that θ has a U-shaped form, while the three other metrics are inverted U-shaped. Interestingly, the standard errors $\mathrm{SE}(\delta)$ and $\mathrm{SE}(\tau)$ approach 0 for δ or τ near to 0 or 1. This is not the case for the rank score metric ρ, for which standard errors for $\rho = 0$ and $\rho = 1$ are larger than 0. Assume that country C1 is low-performing (negative θ value) and country C2 has average performance (θ average of about 0). Then, it can be the case that the latent trait is less precisely assessed for country C1 than for country C2 in the θ metric, but more precisely assessed for country C1 than C2 in one of the three alternative metrics δ, τ or ρ. These statements rely on the somewhat arbitrary choice of the latent trait metric used to quantify differences between countries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

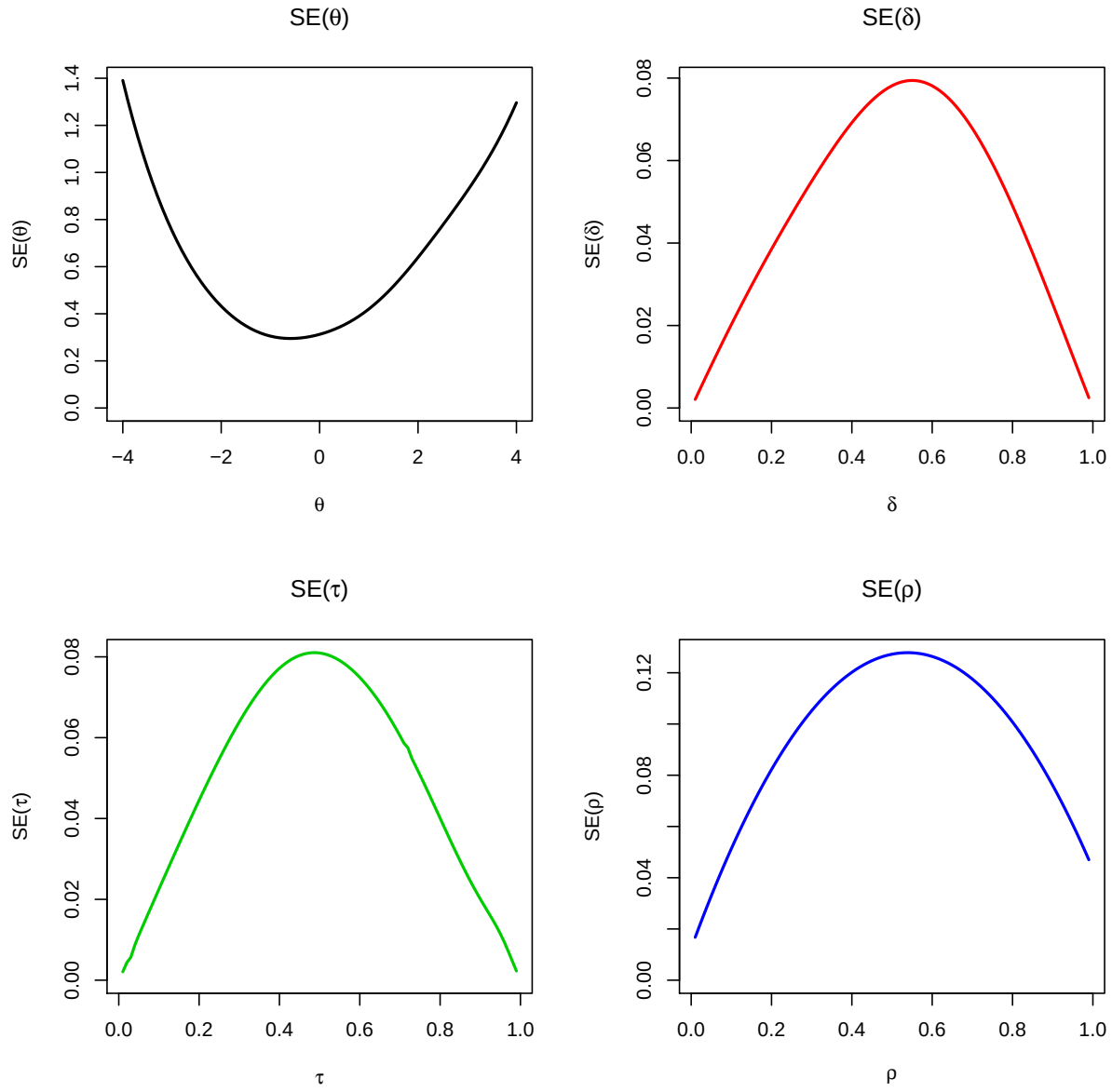

Figure 5. Conditional standard error functions for the logit score θ (upper left figure), the delta score δ (upper right figure), the true score τ (lower left figure), and the rank score ρ (lower right figure) for the PISA 2006 reading test

Table 2 contains detailed results of means, standard deviations, and country ranks based on means for the 26 countries. For the first six high-performing countries, country ranks are the same for all four trait metrics. However, there are countries for which ranks differ considerably. Relatively large deviations are observed for Belgium (BEL; maximum rank difference "maxrk" of 4), Estonia (EST; makrk=6) and Germany (DEU; maxrk=7). The most crucial difference occurs for the τ and the ρ metric. For the three mentioned countries, the standard deviation of θ was relatively low or high compared to all other countries in the sample. This observation explains the differences among ranks because the tails of the θ distributions are differently weighted (i.e., differently transformed) for τ and ρ.

Overall, the Spearman rank correlations of country means ranged between .949 (between τ and ρ) and .992 (between θ and δ). The average rank difference of country means across different metrics was 2.000 (see column "maxrk" in Table 2; SD = 1.853, Min = 0, Max = 7). The Spearman rank correlations of country standard deviations ranged between .973 (between τ and ρ) and .999 (between δ and ρ). The average rank difference of country standard deviations across different metrics was 1.000 (SD = 1.301, Min = 0, Max = 5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

To sum up, the choice of the ability metric can have relevance for some countries for the reporting of country means.

Table 2. Country-level results for PISA 2006 reading for different ability metrics

cnt	Country	N	M				SD				Rank M				
			θ	δ	τ	ρ	θ	δ	τ	ρ	θ	δ	τ	ρ	maxrk
KOR	South Korea	2790	0.471	0.603	0.663	0.646	0.831	0.176	0.166	0.246	1	1	1	1	0
FIN	Finland	2536	0.327	0.576	0.646	0.614	0.570	0.130	0.124	0.193	2	2	2	2	0
CAN	Canada	12142	0.234	0.553	0.616	0.577	0.823	0.179	0.176	0.255	3	3	3	3	0
IRL	Ireland	2468	0.170	0.538	0.599	0.554	0.911	0.193	0.192	0.272	4	4	4	4	0
AUS	Australia	7562	0.144	0.534	0.596	0.550	0.876	0.188	0.189	0.267	5	5	5	5	0
SWE	Sweden	2374	0.098	0.523	0.581	0.535	1.015	0.213	0.214	0.295	6	6	6	6	0
NLD	Netherlands	2666	0.084	0.521	0.577	0.531	1.051	0.219	0.221	0.302	7	7	7	8	1
POL	Poland	2968	0.065	0.515	0.573	0.521	0.981	0.209	0.211	0.293	8	9	8	9	1
BEL	Belgium	4840	0.031	0.517	0.567	0.532	1.278	0.250	0.257	0.333	9	8	11	7	4
JPN	Japan	3203	0.015	0.507	0.562	0.512	1.103	0.225	0.229	0.308	10	10	13	10	3
CHE	Switzerland	6578	0.015	0.506	0.569	0.511	0.852	0.186	0.190	0.265	11	11	10	11	1
DNK	Danmark	2431	0.008	0.502	0.566	0.503	0.828	0.181	0.183	0.260	12	12	12	14	2
EST	Estonia	2630	0.002	0.501	0.571	0.501	0.616	0.142	0.143	0.211	13	13	9	15	6
GBR	Great Britain	7061	−0.028	0.498	0.557	0.500	0.989	0.206	0.211	0.286	14	15	15	16	2
FRA	France	2524	−0.039	0.500	0.559	0.508	1.004	0.206	0.215	0.285	15	14	14	13	2
ISL	Iceland	2010	−0.055	0.489	0.556	0.486	0.741	0.165	0.170	0.239	16	18	16	18	2
AUT	Austria	2646	−0.057	0.493	0.547	0.495	1.125	0.230	0.237	0.314	17	17	17	17	0
DEU	Germany	2701	−0.098	0.497	0.539	0.510	1.485	0.280	0.290	0.364	18	16	19	12	7
HUN	Hungary	2399	−0.110	0.477	0.544	0.468	0.694	0.156	0.163	0.229	19	20	18	20	2
NOR	Norway	2504	−0.135	0.479	0.535	0.478	1.079	0.221	0.231	0.303	20	19	21	19	2
ESP	Spain	10506	−0.168	0.460	0.535	0.440	0.432	0.102	0.108	0.155	21	23	20	23	3
LUX	Luxembourg	2443	−0.210	0.463	0.519	0.456	1.073	0.219	0.231	0.300	22	21	22	22	1
PRT	Portugal	2773	−0.219	0.455	0.517	0.439	0.863	0.185	0.195	0.262	23	24	23	24	1
CZE	Czech Republic	3246	−0.237	0.462	0.506	0.457	1.398	0.270	0.280	0.355	24	22	24	21	3
ITA	Italy	11629	−0.288	0.443	0.502	0.426	0.966	0.199	0.212	0.276	25	25	25	25	0
GRC	Greece	2606	−0.385	0.419	0.479	0.388	0.868	0.183	0.196	0.256	26	26	26	26	0

Note. cnt = country label; N = sample size per country; M = mean; SD = standard deviation; Rank M = country rank with respect to mean M; θ = logit ability metric from two-parameter logistic (2PL) model; δ = metric of the latent D-scoring (LDS) model; τ = true score metric; ρ = rank score metric; maxrk = maximum rank difference among ability metrics θ, δ, τ, and ρ.

5. Discussion

This article shows that the newly proposed LDS model of Dimitrov can be interpreted as a reparametrization of the well-studied 2PL model. Hence, all established statistical techniques for the 2PL model can be used for practical applications of the LDS model. It has been shown that the latent trait score δ from the LDS model is a monotonous (logistic) transformation of the θ score from the 2PL model. All other psychometric areas like differential item functioning, equating and linking, or test assembly must not be reinvented for the LDS model because known techniques for the 2PL model can be used.

Our derivations make it clear that the LDS model with one item parameter is equivalent to the 1PL model. Hence, identification constraints from the 1PL model translate into identification constraints of the LDS model with one item parameter. Dimitrov also proposes an LDS model with three item parameters [16]. In this extension, a guessing parameter appears in the IRF. It is evident that this model is equivalent to the three-parameter logistic IRT model.

Although these findings might be interpreted as somehow destructive for the research surrounding the LDS model, we do not think that the LDS model is not of interest at all. We wanted to argue that the choice latent trait metric is arbitrary in IRT models, and the θ or the δ metric can be both useful in applications. We tend to prefer bounded trait metrics in applications because it seems more challenging to interpret the possibility of unbounded

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

negative and positive trait values of θ [33]. However, we would prefer the true score metric τ or the rank score ρ over δ. The latent D-score δ can be interpreted as a particular true score in which only a reference item with $a_i = 1$ and $b_i = 0$ is used. We believe that using a well-chosen to reference tests with its item parameters provides a better interpretable latent trait metric in practical applications. The rank score ρ has the advantage that it does not depend on item parameters. For example, in the PISA study, one fixes the θ metric in the starting study (e.g., in PISA 2000) to a mean of 500 and a standard deviation of 100. Using the rank metric ρ would imply that the metric is identified by assuming a uniform distribution on $[0, 1]$ for identification. Both approaches might be legitimate in practical applications.

In IRT models, items are typically treated as fixed. However, they can alternatively be interpreted as exchangeable. Item sampling models [34–36] have fewer assumptions in this respect and could be alternatively employed in assessment studies.

The LDS model has been motivated as an IRT analog of the so-called manifest D-scoring method [37]. The scoring rule $\sum_{i=1}^{I}(1 - \pi_i)X_i$ is used in this approach, where $\pi_i = \mathrm{P}(X_i = 1)$ is the probability of getting item i correct. In manifest D-scoring, more difficult items receive larger weights. This property might have appeal in some applications. However, we believe that this scoring rule does not adequately represent all items in a test in typical assessment studies and might lead to country comparisons with reduced validity.

Author Contributions: XXX

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The PISA 2006 dataset is available from https://www.oecd.org/pisa/pisaproducts/database-pisa2006.htm.

Acknowledgments: We would like to thank Dimiter Dimitrov for helpful explanations about motivations of the D-scoring method.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

1PL	One-parameter logistic model
2PL	Two-parameter logistic model
IRF	Item response function
IRT	Item response theory
JML	Joint maximum likelihood
LDS	Latent D-scoring model
MML	Marginal maximum likelihood
PISA	Programme for international student assessment

Appendix A. Item Parameter Estimates for the PISA 2006 Reading Data

In Table A1, estimated item parameters from the 2PL model are shown (columns "$\hat{a}_i$" and "$\hat{b}_i$"). In addition, transformed item parameter for the LDS model are displayed in the columns "$\hat{\alpha}_i$" and "$\hat{\beta}_i$"

In Figure A1, IRFs of the following seven selected items are shown: R067Q01, R104Q02, R104Q05, R111Q02B, R219Q01T, R219Q02, and R220Q01.

 doi:10.20944/preprints202105.0699.v1

Table A1. Estimated item parameters for the PISA 2006 reading dataset

		2PL		LDS	
Item	π_i	a_i	b_i	α_i	β_i
R055Q01	0.817	1.395	−1.486	1.395	0.185
R055Q02	0.480	1.379	0.043	1.379	0.511
R055Q03	0.584	1.620	−0.334	1.620	0.417
R055Q05	0.719	2.118	−0.778	2.118	0.315
R067Q01	0.892	1.227	−2.072	1.227	0.112
R067Q04	0.382	0.832	0.723	0.832	0.673
R067Q05	0.582	1.088	−0.307	1.088	0.424
R102Q04A	0.343	1.460	0.669	1.460	0.661
R102Q05	0.457	1.330	0.244	1.330	0.561
R102Q07	0.842	1.417	−1.493	1.417	0.183
R104Q01	0.816	1.627	−1.322	1.627	0.211
R104Q02	0.326	0.584	1.333	0.584	0.791
R104Q05	0.046	1.132	3.131	1.132	0.958
R111Q01	0.643	1.365	−0.604	1.365	0.353
R111Q02B	0.155	1.046	1.912	1.046	0.871
R111Q06B	0.351	1.588	0.542	1.588	0.632
R219Q01E	0.582	1.633	−0.250	1.633	0.438
R219Q01T	0.699	1.860	−0.664	1.860	0.340
R219Q02	0.792	1.534	−1.179	1.534	0.235
R220Q01	0.434	1.762	0.305	1.762	0.576
R220Q02B	0.621	1.520	−0.376	1.520	0.407
R220Q04	0.596	1.302	−0.312	1.302	0.423
R220Q05	0.823	1.977	−1.145	1.977	0.241
R220Q06	0.669	1.167	−0.675	1.167	0.337
R227Q01	0.521	0.778	−0.151	0.778	0.462
R227Q02T	0.337	0.993	0.793	0.993	0.688
R227Q03	0.546	1.664	−0.183	1.664	0.454
R227Q06	0.706	1.766	−0.777	1.766	0.315

Note. 2PL = two-parameter logistic model; LDS = latent D-scoring model; π_i = proportion correct

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 May 2021 doi:10.20944/preprints202105.0699.v1

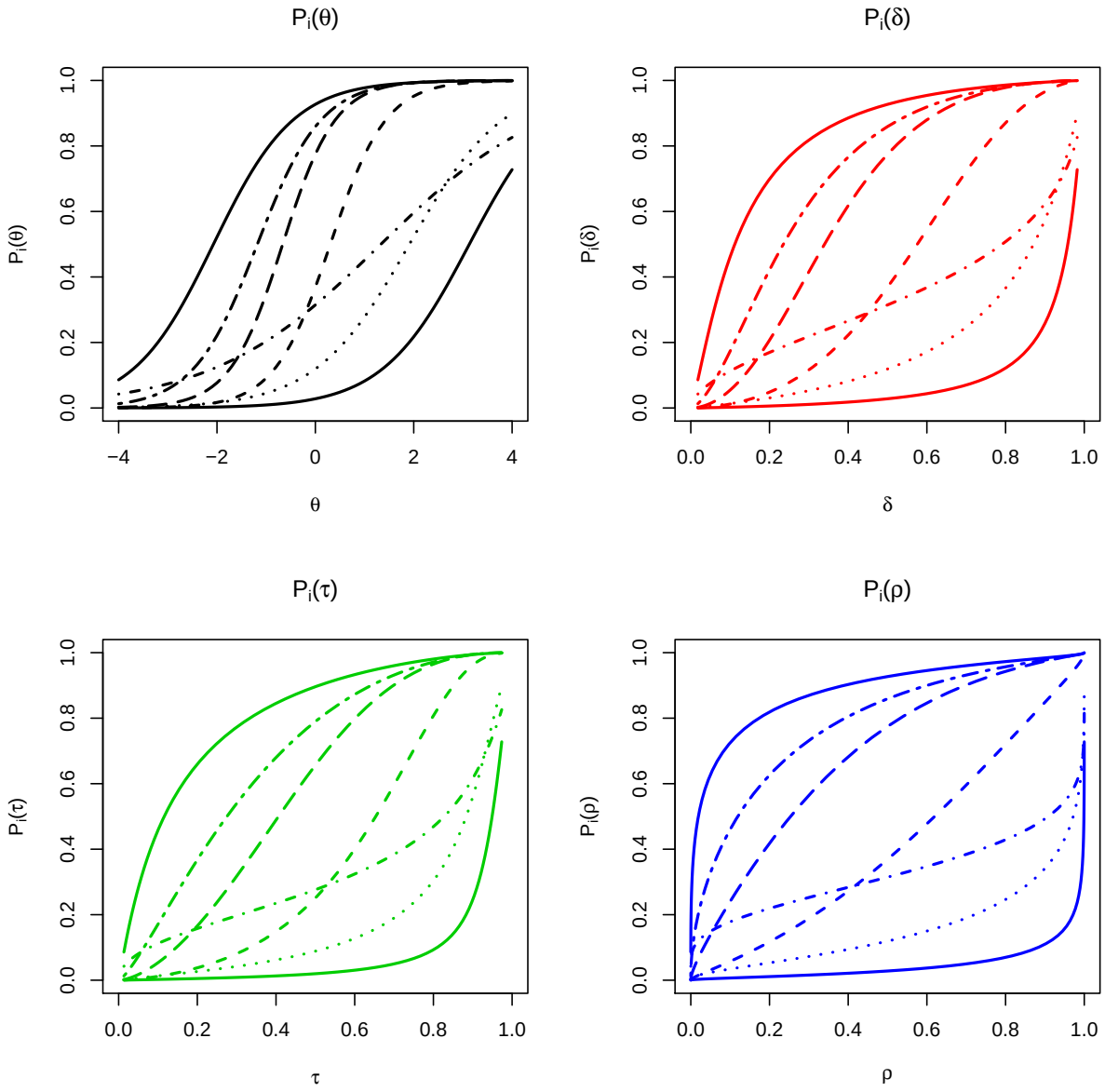

Figure A1. Item response functions of seven selected items from the PISA 2006 reading test

References

1. Baker, F.B.; Kim, S.H. *Item response theory: Parameter estimation techniques*; CRC Press: Boca Raton, 2004. doi:10.1201/9781482276725.
2. Yen, W.M.; Fitzpatrick, A.R. Item response theory. In *Educational measurement*; Brennan, R.L., Ed.; Praeger Publishers, 2006; pp. 111–154.
3. Dimitrov, D.M.; Atanasov, D.V. Latent D-scoring modeling: Estimation of item and person parameters. *Educ. Psychol. Meas.* **2021**, *81*, 388–404. doi:10.1177/0013164420941147.
4. Boos, D.D.; Stefanski, L.A. *Essential statistical inference*; Springer: New York, 2013. doi:10.1007/978-1-4614-4818-1.
5. Ballou, D. Test scaling and value-added measurement. *Educ. Finance Policy* **2009**, *4*, 351–383. doi:10.1162/edfp.2009.4.4.351.
6. Ho, A.D. A nonparametric framework for comparing trends and gaps across tests. *J. Educ. Behav. Stat.* **2009**, *34*, 201–228. doi:10.3102/1076998609332755.
7. Ramsay, J.O. A geometrical approach to item response theory. *Behaviormetrika* **1996**, *23*, 3–16. doi:10.2333/bhmk.23.3.
8. van der Linden, W.J. Unidimensional logistic response models. In *Handbook of item response theory, Volume One: Models*; CRC Press: Boca Raton, 2016; pp. 11–30. doi:10.1201/9781315374512-3.
9. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores*; Lord, F.M.; Novick, M.R., Eds.; MIT Press: Reading, MA, 1968; pp. 397–479.

10. Rasch, G. *Probabilistic models for some intelligence and attainment tests*; Danish Institute for Educational Research: Copenhagen, 1960.
11. Culpepper, S.A. The prevalence and implications of slipping on low-stakes, large-scale assessments. *J. Educ. Behav. Stat.* **2017**, *42*, 706–725. doi:10.3102/1076998617705653.
12. Formann, A.K. Constrained latent class models: Theory and applications. *Brit. J. Math. Stat. Psychol.* **1985**, *38*, 87–111. doi:10.1111/j.2044-8317.1985.tb00818.x.
13. Xu, X.; von Davier, M. *Fitting the structured general diagnostic model to NAEP data.* (Research Report No. RR-08-28). Educational Testing Service, 2008. doi:10.1002/j.2333-8504.2008.tb02113.x.
14. Brennan, R.L. Misconceptions at the intersection of measurement theory and practice. *Educ. Meas.* **1998**, *17*, 5–9. doi:10.1111/j.1745-3992.1998.tb00615.x.
15. Dimitrov, D.M.; Atanasov, D.V. Testing for differential item functioning under the D-scoring method. *Educ. Psychol. Meas.* **2021**. [Epub ahead of print], doi:10.1177/00131644211001524.
16. Dimitrov, D.M. Modeling of item response functions under the D-scoring method. *Educ. Psychol. Meas.* **2020**, *80*, 126–144. doi:10.1177/0013164419854176.
17. Dimitrov, D.M.; Atanasov, D.V. An approach to test equating under the latent D-scoring method. *Meas. Interdiscip. Res. Persp.* **2021**. [In press].
18. Han, K.C.T.; Dimitrov, D.M.; Al-Mashary, F. Developing multistage tests using D-scoring method. *Educ. Psychol. Meas.* **2019**, *79*, 988–1008. doi:10.1177/0013164419841428.
19. Hoff, P.D. *A first course in Bayesian statistical methods*; Springer: New York, 2009. doi:10.1007/978-0-387-92407-6.
20. Atchison, J.; Shen, S.M. Logistic-normal distributions: Some properties and uses. *Biometrika* **1980**, *67*, 261–272. doi:10.1093/biomet/67.2.261.
21. R Core Team. *R: A language and environment for statistical computing*, 2020. Vienna, Austria. https://www.R-project.org/.
22. Robitzsch, A. *sirt: Supplementary item response theory models*, 2020. R package version 3.9-4. https://CRAN.R-project.org/package=sirt.
23. von Davier, M. A general diagnostic model applied to language testing data. *Brit. J. Math. Stat. Psychol.* **2008**, *61*, 287–307. doi:10.1348/000711007X193957.
24. Aitkin, M. Expectation maximization algorithm and extensions. In *Handbook of item response theory, Vol. 2: Statistical tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, 2016; pp. 217–236. doi:10.1201/b19166-12.
25. Magis, D. A note on the equivalence between observed and expected information functions with polytomous IRT models. *J. Educ. Behav. Stat.* **2015**, *40*, 96–105. doi:10.3102/1076998614558122.
26. Brennan, R.L. Perspectives on the evolution and future of educational measurement. In *Educational measurement*; Brennan, R.L., Ed.; Praeger Publishers, 2006; pp. 1–16.
27. Yamamoto, K.; Shin, H.J.; Khorramdel, L. Multistage adaptive testing design in international large-scale assessments. *Educ. Meas.* **2018**, *37*, 16–27. doi:10.1111/emip.12226.
28. Bonifay, W. *Multidimensional item response theory*; Sage: Thousand Oaks, 2019.
29. Reckase, M.D. *Multidimensional item response theory models*; Springer: New York, 2009. doi:10.1007/978-0-387-89976-3.
30. OECD. *PISA 2006. Technical report*; OECD: Paris, 2009.
31. Oliveri, M.E.; von Davier, M. Analyzing invariance of item parameters used to estimate trends in international large-scale assessments. In *Test fairness in the new generation of large-scale assessment*; Jiao, H.; Lissitz, R.W., Eds.; Information Age Publishing: New York, 2017; pp. 121–146.
32. Robitzsch, A. Robust Haebara linking for many groups: Performance in the case of uniform DIF. *Psych* **2020**, *2*, 155–173. doi:10.3390/psych2030014.
33. Ramsay, J.O.; Li, J.; Wiberg, M. Better rating scale scores with information–based psychometrics. *Psych* **2020**, *2*, 347–369. doi:10.3390/psych2040026.
34. van der Linden, W.J. Binomial test models and item difficulty. *Appl. Psychol. Meas.* **1979**, *3*, 401–411. doi:10.1177/014662167900300311.
35. Wiley, J.A.; Martin, J.L.; Herschkorn, S.J.; Bond, J. A new extension of the binomial error model for responses to items of varying difficulty in educational testing and attitude surveys. *PLOS ONE* **2015**, *10*, e0141981. doi:10.1371/journal.pone.0141981.
36. Hong, H.; Wang, C.; Lim, Y.S.; Douglas, J. Efficient models for cognitive diagnosis with continuous and mixed-type latent variables. *Appl. Psychol. Meas.* **2015**, *39*, 31–43. doi:10.1177/0146621614524981.
37. Dimitrov, D.M. An approach to scoring and equating tests with binary items: Piloting with large-scale assessments. *Educ. Psychol. Meas.* **2016**, *76*, 954–975. doi:10.1177/0013164416631100.