Large-scale Assessments in Education

RESEARCH Open Access

Generating group-level scores under response accuracy-time conditional dependence

Hyo Jeong Shin[1*], Paul A. Jewsbury[1] and Peter W. van Rijn[2]

*Correspondence: hshin@ets.org

[1] Educational Testing Service, 660 Rosedale Road, 13-E, Princeton, NJ 08541, USA
[2] ETS Global, Amsterdam, The Netherlands

Abstract

The present paper investigates and examines the conditional dependencies between cognitive responses (RA; Response Accuracy) and process data, in particular, response times (RT) in large-scale educational assessments. Using two prominent large-scale assessments, NAEP and PISA, we examined the RA-RT conditional dependencies within each item in the measurement model and the structural model. Evidence for RA-RT conditional dependencies was observed in data from both programs, presenting a challenge in incorporating RT to the current operational models in NAEP and PISA that do not account for RA-RT conditional dependencies. However, inclusion of RT in the model had a relatively large contribution to improving the measurement of ability (residual variance decrease of 11% in NAEP and 18% in PISA), in contrast to relatively modest difference in parameter estimation from neglecting the conditional dependencies (e.g., estimated difference on residual variance of 1% in both NAEP and PISA). We conclude that the benefits of incorporating RT in the operational models for large-scale educational assessments may outweigh the costs.

Keywords: Response time, Big data in education, Conditional dependence, Large-scale assessments, NAEP, PISA

Introduction

In many large-scale assessments (LSAs), digital-based testing has become the primary mode of administration. For example, the Programme for International Student Assessment (PISA) transitioned from a paper-based to a computer-based assessment (CBA) in the 2015 cycle, and the Programme for the International Assessment of Adult Competencies (PIAAC) transitioned to CBA in 2012. Another prominent LSA program, the United States' National Assessment of Educational Progress (NAEP), first administered a CBA in 2011 and transitioned the four main assessments of reading and mathematics at grades 4 and 8 to CBA in 2017 (NCES, 2018a, 2018b) . Notably, CBA enables the collection of *process data*, which may include response times (RT), number of actions, and keystrokes. PISA started providing process data variables, in addition to traditional item responses, in the Public Use File (PUF; OECD, 2017, 2020) starting with the 2015 cycle, and NAEP also provides RT and process data upon request.

Springer Open

The availability of process data collected in LSAs has led to considerable research on a variety of topics. Process data can be useful for improving data quality, test security, reliability, and validity (Bergner and von Davier, 2019; Ercikan et al., 2017); von Davier et al., 2019). For example, process data can be used to evaluate the validity of cognitive responses and whether the test has been administered in compliance with the technical standards Yamamoto and Lennon (2018), or to provide insights on the mechanisms of non-responses in low-stakes assessments and how the missing data should be treated in the analysis (e.g., Lu & Wang, 2020; Pohl et al., 2019, Ulitzsch et al., 2020b; Weeks et. al., 2016). In particular, such non-response behavior unveiled by process data can be useful for investigating student effort and engagement (e.g., Ulitzsch et al., 2020a; Michaelides et al., 2020). A recent study by Pohl, Ulitzsch, and von Davier (2021) argued that disentangling and reporting test-taking behaviors based on process data can improve the comparability and interpretability of the reported scores. For example, test takers from some countries may perform better than test takers from other countries by taking more time and responding to fewer tasks. Such an insight can provide rich and relevant information for policy makers. Furthermore, classifying respondents into classes defined by process data contributes to deepening our understanding of test-taking behaviors and student inquiry (e.g., Greiff et al., 2015; He et al., 2018; Teig et al., 2020).

In many LSAs, a primary goal is to provide accurate group-level estimates that allow for comparisons over time to measure trends (Kirsch et al., 2013). An important question then is how newly available process data can be utilized to improve group-level reporting without threatening the comparability of scores over time. Yet, to our knowledge, most research using LSA process data has focused on relatively small subsets of items and sub-samples of students who provided data to those subsets of items (von Davier et al., 2019). In-depth investigation with limited data can be informative to contextualize test-taking behaviors, but cannot guide methodological advances in generating reporting group-level scores in LSAs due to the limited use of data and limited implications for reporting of scores. LSAs, such as NAEP and PISA, involve item response theory-latent regression models (IRT-LRMs) to generate plausible values (PVs) for every student, which are used for both the official results and for secondary analysis.[1] PVs are multiply imputed values (Rubin, 1996) for latent proficiencies that enable secondary users to examine the relationships between proficiencies and contextual variables with simple statistical models commonly available in statistical packages, such as regression models and t-tests. Analysis with multiply imputed values must be congenial with the imputation model to avoid a source of estimation bias (Meng, 1994). In the context of LSAs, this means that if the relationship between ability and a given contextual variable is of interest to secondary users, that contextual variable must be included as a covariate in the IRT-LRM. This is the main reason why hundreds of covariates are often included in the IRT-LRM to accommodate as many potential secondary-user analyses as possible (von Davier et. al. 2006). Analogously, process data must be incorporated in the PV generation to ensure correct inferences about the relationships between proficiencies and process data.[1]

[1] Details about analytic procedures used in LSA, such as NAEP and PISA, are well-described in Mislevy et. al. (1992), von Davier and Sinharay (2013); von Davier et. al. (2006); von Davier et. al. (2009).

As a consequence, PISA 2018 cycle started to incorporate the process data (more precisely, RT) in the PV generation process (OECD, 2020; Shin et al., in press). In this approach, student-level RT were pre-processed to be included in the IRT-LRM as a predictor in addition to the other covariates. However, the degree to which incorporating process data in the PV-generating procedures violates the assumptions underlying the IRT-LRM is largely unknown, and gathering further insights on how to best incorporate process data in the IRT-LRM is important. IRT-LRMs in LSAs follow the standard assumption of independence between the cognitive item responses and regression covariates, conditional on latent ability (Mislevy, 1985; von Davier et al., 2006). This conditional independence assumption is typically reasonable for contextual variables that are measured independently from performance on the test (response accuracy; RA), such as demographics and background questionnaire responses (or evaluated with measurement non-invariance methods during routine operational procedures; Meredith 1993). However, RT and other process data are not measured independently of performance on the test, and some degree of conditional dependence can therefore be expected (Bolsinova et al., 2017).

In this paper, we examine the impacts of RA-RT conditional dependence, which violates a central assumption in utilizing process data in the statistical analyses used in LSAs. While the literature on RA-RT conditional dependencies is extensive (for reviews, see Bolsinova et al., 2017, De Boeck & Jeon, 2019), the magnitude and impact of such conditional dependencies has, to our knowledge, not been previously evaluated within the context of LSAs. In this paper, we examine the issue of RA-RT conditional dependence within the context of LSAs. First, we estimate the significance and magnitude of the conditional dependencies in LSAs by applying methods from the literature to NAEP and PISA data in our first research question (RQ1). Second, we evaluate the impact on parameter estimation of applying the current LSA operational methods, which assume conditional independence, to data that has conditional dependencies in research questions 2 and 3 (RQ2 and RQ3).

- RQ1: What is the degree to which RA-RT conditional dependencies are found in NAEP and PISA data?
- RQ2: What are the empirical consequences of including RT in the IRT-LRMs?
- RQ3: What are the empirical consequences of ignoring conditional dependencies in the IRT-LRMs?

Conditional independence assumptions in the IRT-LRM

Fundamental assumptions in the IRT-LRM

An IRT-LRM consists of two components: a measurement model (the item response theory model; IRT model) and a structural model (the latent regression model; LRM). Two types of observed variables can be distinguished: measurement variables $\mathbf{x}$ (the item responses from the cognitive assessment that are assumed to measure the latent ability in the IRT model) and contextual variables (other variables collected about the students that are assumed to relate to the latent ability, such as gender and parental education). As thousands of contextual variables may be collected, the standard operational practice is to reduce the dimensionality of the contextual variables by means of principal component analysis (PCA), and a selected set of principal components is used as covariates or predictors in the LRM, **z**. With x_j denoting the response to item j in the test, two primary assumptions of IRT-LRM are:

- *Assumption 1.* Conditional independence between item responses given the latent ability:

$$P(\mathbf{x}|\theta) = \prod_j P(x_j|\theta). \tag{1}$$

- *Assumption 2.* Conditional independence between item responses and contextual variables given the latent ability:

$$P(\mathbf{x}|\theta, \mathbf{z}) = P(\mathbf{x}|\theta). \tag{2}$$

With these two assumptions, we can find the conditional density of the latent ability as follows:

$$f(\theta|\mathbf{x}, \mathbf{z}) = \frac{P(\mathbf{x}|\theta)f(\theta|\mathbf{z})}{\int P(\mathbf{x}|\theta)f(\theta|\mathbf{z})\,d\theta}, \tag{3}$$

where $f(\theta|\mathbf{z})$ is usually assumed to be a normal distribution. This conditional density plays a seminal role in the analysis of LSAs (e.g., Mislevy et al., 1992; von Davier et al., 2006, 2009).

Additional assumptions associated with process data

There are at least two possible ways to include RT and other process data in the IRT-LRM: in the measurement model along with item responses or in the structural model as predictors. Entering RT or process data as measurement variables by combining RT in the scoring rubric (van Rijn and Ali, 2018a, 2018b), or by utilizing as collateral information (Bolsinova and Tijmstra, 2018a; Reis Costa et al., 2021), proved to be promising to improve the precision of ability estimation.

In the measurement model, a useful starting point for including RT and studying conditional dependencies is the hierarchical model (van der Linden, 2007) in which an IRT model is used for the item responses and a log-normal factor model for RT. Based on the hierarchical model, two additional conditional independence assumptions are imposed. Note that t_j indicates the RT for item j, and τ indicates a latent speed variable.

- *Assumption 3.* Conditional independence among RTs given latent speed:

$$f(\mathbf{t}|\tau) = \prod_j f(t_j|\tau). \tag{4}$$

- *Assumption 4.* Conditional independence between RTs and responses given latent ability and latent speed:

$$f(\mathbf{x}, \mathbf{t}|\theta, \tau) = P(\mathbf{x}|\theta, \tau)f(\mathbf{t}|\theta, \tau), \tag{5}$$

where $\mathbf{t}$ is the vector of RTs for all items. Glas (2010) elaborated three different assumptions of conditional independence for the hierarchical framework, and each of those assumptions corresponds to *Assumptions 1, 3* and *4*.

Finally, as the most relevant assumption to this paper, in order to include RT in the IRT-LRM based on van der Linden's, (2007) hierarchical model, we also need the following conditional independence assumption for RTs and contextual variables.

- *Assumption 5.* Conditional independence between RTs and contextual variables given latent speed:

$$f(\mathbf{t}|\tau, \mathbf{z}) = f(\mathbf{t}|\tau). \tag{6}$$

Now, the conditional density of the latent variables on the observed variables becomes

$$f(\theta, \tau|\mathbf{x}, \mathbf{t}, \mathbf{z}) = \frac{P(\mathbf{x}|\theta)f(\mathbf{t}|\tau)f(\theta, \tau|\mathbf{z})}{\int P(\mathbf{x}|\theta)f(\mathbf{t}|\tau)f(\theta, \tau|\mathbf{z})\,d\theta\,d\tau}. \tag{7}$$

Our paper deals with RT variables since these are more commonly studied and considered as important process data features, but conditional independence can be studied for other types of process data as well (e.g., number of actions) using the same principles outlined here.

Testing conditional independence

A comprehensive overview of psychometric models for RT and RA is provided in De Boeck and Jeon (2019). In their review, the most relevant psychometric model to our study is categorized as *local dependency models*. This approach includes models in which RA and RT are jointly modeled but in which extra dependency is modeled beyond the relationship of their latent variables and item parameters. Within this approach, they further categorized two types of models. The first type includes latent variable models with remaining dependencies through the local dependency parameters. More specifically, local dependency parameters can be incorporated into the model as residual correlations (e.g., Ranger & Ortner, 2012) or the direct effect of RT on the corresponding RA (e.g., Bolsinova et al., 2017, De Boeck et al., 2017). The second type includes models that classify the item responses based on different response mechanisms. Classes can be defined through the observed RT variables, such as a fast mode and a slow mode (e.g., Partchev & De Boeck, 2012), or latent classes can be estimated (e.g., Molenaar & de Boeck, 2018, Wang & Xu, 2015). De Boeck (2017) also provide a thorough overview of joint models for RA and RT focusing on modeling conditional dependence. They suggest the possibility to move toward explanatory models to better understand response processes that may lead to conditional dependencies.

In this paper, we focus on *parametric* methods in which conditional independence assumptions are tested through the estimation of parameters in existing or newly proposed psychometric models with relatively strong distributional assumptions. Within the hierarchical model framework, van der Linden and Glas (2010) proposed Lagrange multiplier tests for evaluating three conditional independence assumptions (*Assumptions 1, 3* and *4*) when item parameters are known, and Glas and van der Linden (2010) further derived analogous tests when all structural parameters are estimated using marginal maximum likelihood estimation. Alternatively, Molenaar, Tuerlinckx, and van der Maas Molenaar et. al. (2015) developed a bivariate generalized linear IRT approach in which separate generalized linear measurement models for responses and for RT variables are subsequently linked by cross-relations. Because of the flexibility of the model, this approach can be used to test those three aforementioned assumptions. In addition, this generalized modeling framework is further advantageous to model the residual correlations for items that exhibit conditional dependencies. Modelling residual correlations to relax the conditional independence assumption was also proposed in Ranger and Ortner (2012), focusing on residual correlations between RT and RA for within items. In this model, additional item-specific parameters reflect the remaining within-item RA-RT relationship that is not explained by the latent correlation between ability and speed (see Analysis section below for equations). That is, conditional dependencies are allowed to vary across items. Later Meng et. al. (2015) further extended this model by allowing conditional dependencies to vary across items as well as persons. More formally, Ranger and Ortner's (2012) model corresponds to allowing possible effects of RT on the intercept of the item response function (Bolsinova et al., 2017), and Bolsinova et. al. (2017) proposed an extended model that incorporates the effects of the residual RT on the slope as well as the intercept parameter. The latter can capture differences in item discrimination for slow and fast responses, which is a pattern found by, for example, Goldhammer et. al. (2014). In this case, the probability of a correct item response interacts with the item difficulty, as well as its expected speed. More recently, relaxing the monotone and linear conditional dependence was attempted through modeling a nonlinear, such as quadratic or multiple-category, conditional dependence (Bolsinova and Molenaar, 2018).

Alternative to the parametric methods reviewed above, approaches with fewer or weaker assumptions can be employed as well. For example, posterior predictive checks can be used for testing RA-RT conditional dependencies (Bolsinova and Tijmstra, 2016), as a flexible method with respect to which model is chosen. Furthermore, Bolsinova and Maris (2016) proposed a *non-parametric* approach for testing conditional independence between RA and RT based on the Kolmogorov-Smirnov (KS) test. Another relevant approach for jointly modeling RA and RT can be found in van Rijn and Ali (2018a) where RT information is utilized in the scoring rule: slower responses, regardless of whether they are correct or incorrect, contribute less to the item score. De Boeck and Jeon (2019) involved this generalized speed-accuracy response model as one of the local dependency models because being correct and fast or slow are combined in the scoring rules, although RA-RT conditional dependence is not explicitly modeled.

As the literature review shows, RA-RT conditional dependencies prevail in most of the cognitive assessments, and various statistical methods have proven useful to test and understand response processes. For instance, positive conditional dependence is interpreted as the unexplained association between slower and more accurate responses, while negative conditional dependence as the unexplained association between faster and more accurate responses. In this study, we aim to contribute to the literature in two ways by focusing on conditional dependencies in the context of LSAs. First, we empirically evaluate RA and RT dependencies with NAEP and PISA datasets to examine whether conditional dependencies are found in LSA data. Second, we evaluate the impact on the operational IRT-LRMs as used in LSAs. Although we selected RT, our analysis is critical to other types of process data obtained from the cognitive assessment in LSAs as well.

Methods

Data

In this study, we analyzed the data from the NAEP 2017 mathematics and PISA 2015 science assessments. The domains of mathematics for NAEP and science for PISA were chosen as the largest available datasets. Analyzed data for NAEP included about 126,700 students, involving 178 items in total. In PISA 2015, science was the major domain, taken by all participating students. However, all students are assessed on at least one other domain as well (e.g., mathematics, reading), which is not the case in NAEP. For PISA, we used data collected in English from the United States and Canada to have sufficient sample size for the analyses, comprising about 24,600 student responses on 183 science items. Both NAEP and PISA data consisted of a mixture of multiple-choice items and open-ended response items.

In NAEP 2017 and PISA 2015, data were collected through a balanced incomplete block (BIB) design (Messick et al., 1983), where only a subset of the complete item pool was administered to each student. All items were assigned to one 30-minute *item block* (or *clusters* in PISA terminology). Each student received only two item blocks out of twelve in PISA, and two 30-min blocks out of ten in NAEP. Each item block was presented in approximately the same proportions in the first and second half of the assessments, following the *balanced* block design generally used in LSAs (Mazzeo and von Davier, 2008). Note that because each student receives a small proportion of items in the total item pool, the datasets have a massive amount of missing data.

Concerning the test administration, each item block has a separate time limit in NAEP, while two blocks are sequentially taken with a shared time limit of one hour in PISA (e.g., students can spend more than 30 minutes on the items of first block provided they spend less than 30 minutes on the second block). To minimize confounding from time limits and to handle the massive missing problem, each separately timed section (or *timing section*) was analyzed as a separate dataset. Specifically, each NAEP block was treated as a separate dataset for analysis, while each possible pairing of PISA blocks was treated as a separate dataset. The pairing of PISA blocks ignored the order of the two blocks to ensure sufficient sample sizes for each analysis (e.g., students who took 'S07' and 'S08' item blocks were grouped together regardless of the order, 'S07' then 'S08' or 'S08' then 'S07').[2]

Such data handling resulted in 10 timing sections for NAEP 2017 and 36 timing sections for PISA 2015 data. Item blocks in the NAEP timing sections were mutually exclusive, while each student participated in two timing sections. In contrast, PISA item blocks were not mutually exclusive across timing sections, but no students appeared more than once in different timing sections. In the PISA data, two timing sections turned out to have sparse data due to the high frequencies for certain categories of

[2] We acknowledge that ignoring the order of blocks may affect conditional dependencies *between* items but this is not the focus of our analysis.

Table 1 Distribution of sample size and number of items per timing section, and the list of contextual variables

	PISA				**NAEP**			
	Mean	**SD**	**Min.**	**Max.**	**Mean**	**SD**	**Min.**	**Max.**
Number of students	677	29	610	744	25305	300	24935	25861
Number of items	31	3	25	35	18	3	14	21
Contextual variables (**z**)	HISCED				LEP		SCHTYPE	
	GENDER				IEP		SLUNCH1	
	IMMIG				CENSREG		SRAC10E	
	NBOOKS				DSEX		PARED	

certain contextual variables. Analyses could not be completed successfully in such cases, thus, 34 groups are reported in this paper except those two groups.[3]

Table 1 presents the distributions of sample size and the number of administered items per timing section. Furthermore, contextual variables were also listed. Because our paper pertains to the impact of conditional dependencies on the estimation of IRT-LRM, some of our modeling choices (M3, M4, and M5 in Fig. 2) included the regression of ability on contextual variables. Due to the computational complexity of the full NAEP and PISA IRT-LRMs that involve the regression of the latent variable on hundreds or thousands of variables, we approximate the full IRT-LRMs by including only key reporting variables in the regression as a way to compare the models in a relative manner to evaluate the impact of conditional dependencies. The selected key reporting variables that were used as covariates in the NAEP and PISA regressions are listed in Table 1, and descriptions of those variables are provided in the Appendix.

Analysis

To address the three research questions, five models were analyzed, including two measurement models (M1 and M2) and three latent regression models (M3, M4, and M5). To address our first research question, we focused on directly evaluating the magnitude of conditional dependencies using M1 and M2 in a way that is comparable to the literature. To adress research questions two and three, we focused on evaluating the impact of conditional dependencies using M3, M4, and M5 on the operational method in LSAs, the IRT-latent regression, involving speed as a predictor of ability. To evaluate the dependencies in the context of the measurement model, we employed a parametric method involving parameters for within-item RA-RT dependencies (Ranger and Ortner, 2012) as a generalization of the hierarchical model (van der Linden, 2007). To evaluate the dependencies in the context of the structural model, we extended the IRT-LRM to account for within-item conditional dependencies following a similar approach as used in the measurement model evaluation.

For M1, we start with the extended version of the hierarchical model, allowing item-specific time discrimination parameters (Molenaar et al., 2015), in which conditional

[3] There are two sets of timing variables published for PISA 2015. In this study, we used the updated timing variables posted in November 2020 for PISA 2015 (http://www.oecd.org/pisa/data/2015database/). These updated timing variables report total time spent on the given item by summing over multiple item visits, if applicable.

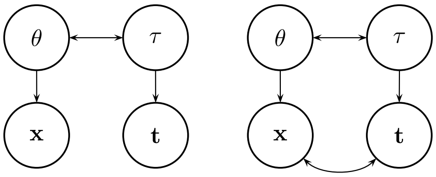

Fig. 1 Measurement Models: M1 (left) and M2 (right)

independence between RA and RT is assumed. Similarly, Reis Costa et. al. (2021 and Bolsinova and Tijmstra (2018) considered the original hierarchical model as the baseline model in their studies for including RT variables in the measurement model. Building on this, we specify additional parameters for modeling within-item RA-RT conditional dependencies for M2. M2 is most relevant to test *Assumption 4*, and a comparison between M1 and M2 informs the presence of within-item RA-RT conditional dependencies in the measurement model empirically observed in the NAEP 2017 and PISA 2015 data. Moving to the structural model, M3 is considered a baseline model, which is a simplified version of IRT-LRM specifying a limited number of regression predictors as listed in Table 1. M4 is an extended IRT-LRM that incorporates RT variables into M3, while maintaining the assumption of conditional independence. The most general model in our analysis is M5 in which within-item RA-RT conditional dependencies for each item are allowed.

- Model 1 (M1): Hierarchical model for accuracy and speed

The hierarchical model developed by van der Linden (2007) has latent ability and speed variables for RT and for RA, respectively, with a latent correlation between them. Assuming dichotomously scored items, IRT models such as the Rasch, two-parameter-logistic (2PL), or three-parameter-logistic (3PL) model are typically used for the RA part. In this paper, however, we use the two-parameter normal ogive model (2PNO; Lord (1952)) for dichotomously scored items and the generalized 2PNO for polytomously scored items. Using the probit function $\Phi^{-1}(.)$, the 2PNO is given by

$$P(x_j = 1|\theta, b_j, a_j) = \Phi(a_j\theta + bj), \tag{8}$$

where a_j is the item slope parameter, θ is the latent ability variable, and b_j is the item intercept parameter. For the RT part, we use a log-normal factor model, which is extended to have item-specific time discrimination parameters (Molenaar et al., 2015). Denoting log-transformed RT as t_j^*, the model is expressed as

$$t_j^* \sim \mathcal{N}(\beta_j - \gamma_j\tau, \alpha_j^2), \tag{9}$$

where β_j is the item time intensity parameter, γ_j is the item time discrimination parameter, τ is the latent speed variable, and α_j^2 is the residual variance of the log RT for item j. A graphical representation of this model is shown in the left panel of Fig. 1 where the possible relationships between item responses ($\mathbf{x}$) and RT variables ($\mathbf{t}$) are fully captured through the correlation (ρ) between θ and τ.

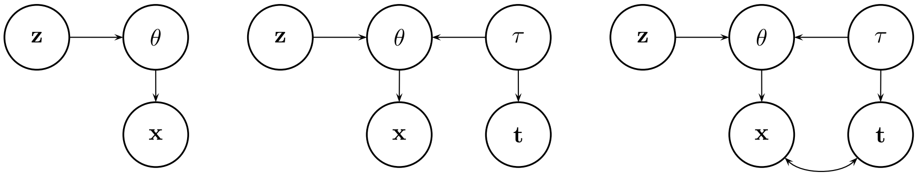

Fig. 2 IRT-LRMs: M3 (left), M4 (middle), and M5 (right)

The two equations above present RA and RT components of the hierarchical model separately. As a way to illustrate the within-item conditional independence assumption, we can express the joint distribution of the underlying continuous RA variable (x_j^*) and log-transformed RT variable (t_j^*) for item j as follows:

$$\begin{bmatrix} x_j^* \\ t_j^* \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} a_j\theta + bj \\ \beta_j - \gamma_j\tau \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \alpha_j^2 \end{pmatrix}\right] \tag{10}$$

- Model 2 (M2): M1 + within-item RA-RT dependencies

As an extension of M1, M2 allows conditional dependencies between RA and RT that can vary for each item. The conditional dependence is expressed as the arrow connecting item responses (**x**) and RT variables (**t**) in the right panel of Fig. 1. The parametric approach proposed in Ranger and Ortner (2012) allows the residuals to be correlated by means of an additional item parameter (π_j). Building on Equation (10), this can be now expressed as:

$$\begin{bmatrix} x_j^* \\ t_j^* \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} a_j\theta + bj \\ \beta_j - \gamma_j\tau \end{pmatrix}, \begin{pmatrix} 1 & \pi_j\alpha_j \\ \pi_j\alpha_j & \alpha_j^2 \end{pmatrix}\right] \tag{11}$$

Note that the off-diagonal elements changed from 0 to $\pi_j\alpha_j$, thus reflecting the item-level parameterization of conditional dependencies. In relation to RQ1, to examine the presence of RA-RT conditional dependencies (*Assumption 4*), the likelihood ratio test that compared M1 (π_j fixed to zero) and M2 (π_j freely estimated) was conducted for each timing section. We employed the robust weighted least squares (WLSMV) method using Mplus 8.0 (Muthén and Muthén, 2017). For the WLSMV method, the conventional approach involving the difference between the chi-square values is not appropriate because the chi-square difference does not follow a chi-square distribution under WLSMV (Muthén and Muthén, 2017). Therefore, we applied the DIFFTEST function that is afforded in Mplus to obtain a correct chi-square difference test as recommended by Muthén and Muthén (2017).

- Model 3 (M3): Baseline IRT-LRM specifying the main contextual variables

M3 is the baseline model for the IRT-LRMs, concerning the latent ability (θ) measured by item responses (**x**) but not latent speed (τ). As seen in the left panel of Fig. 2, the latent ability is regressed on the contextual variables (**z**):

$$\begin{aligned} x_j^* &\sim \mathcal{N}(a_j\theta + bj, 1), \\ \theta \mid \mathbf{z} &\sim \mathcal{N}(\boldsymbol{\Gamma}'\mathbf{z}, \sigma^2) \end{aligned} \tag{12}$$

where $\boldsymbol{\Gamma}$ is a vector of regression coefficients, and σ^2 is the residual variance. In NAEP and PISA operational procedures, the conditional distribution $f(\theta \mid \mathbf{z})$ is typically assumed to be the multivariate normal distribution, allowing correlations among multiple latent abilities (e.g., correlations among math, reading, and science or correlations among subdomains of mathematics). In addition, item parameters (i.e., arrow from θ to item responses, $\mathbf{x}$) are estimated and fixed in the preceding IRT scaling stage so that only the regression coefficients (i.e., arrow from contextual variables $\mathbf{z}$ to θ) and the residual variance σ^2 are estimated to reduce the computational burden (often called as "divide-and-conquer"; Patz & Junker, 1999). In this study, we estimate item parameters and regression coefficients simultaneously, focusing on the case of unidimensional latent ability (θ). Most importantly, note that in M3, ability is regressed on the contextual variables only, and not on the latent speed dimension. This approach is in line with the reporting practices in NAEP and PISA, and we consider M3 as the benchmark structural model.

- Model 4 (M4): M3 + latent speed as an additional predictor

M4 can be viewed as an extension of M3 by specifying the additional predictor of latent speed (τ) measured through observed RT variables. At the same time, this model is an extension of M1 by specifying the contextual variables in the regression to explain latent ability (θ). To aid interpretation, double-sided arrow is used in M1 and M2 to indicate that speed correlates with ability in the measurement model, while a one-sided arrow indicates that speed is a predictor of ability in the IRT-LRM (structural model; M4 and M5). Building on M3, with the use of RT variables, M4 can be expressed as:

$$\begin{aligned} \begin{bmatrix} x_j^* \\ t_j^* \end{bmatrix} &\sim \mathcal{N}\left[\begin{pmatrix} a_j\theta + bj \\ \beta_j - \gamma_j\tau \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \alpha_j^2 \end{pmatrix}\right], \\ \theta \mid \mathbf{z}, \tau &\sim \mathcal{N}(\boldsymbol{\Gamma}'\mathbf{z} + \omega\tau, \sigma^2). \end{aligned} \tag{13}$$

where ω represents the main effect of the latent speed on the latent ability. In particular, as our study is within the context of LSAs and the operational statistical models use these methods, the residual variance (σ^2) is critical in influencing the student conditional posterior distributions, and by turn the within-person variance between PVs, which indicates measurement error. This is consistent with how measurement precision within the operational LSA methods are considered in previous studies (Mislevy, 1991; von Davier et al., 2009; von Davier and Sinharay, 2013). Therefore, the reduction in the estimates of σ^2 between M3 and M4 is interpreted as the measurement precision, which is addressed in RQ2. Having additional predictor should reduce the residual variance unless their correlation is zero, but the extent to which the variance is reduced with the inclusion of RT is unknown.

- Model 5 (M5): M4 + within-item RA-RT dependencies

Unlike general contextual variables, such as gender or parental education, RT variables are co-measured with RA and reflect item-specific processes. This implies that the RA and RT for a given item may be conditionally dependent (given the latent ability). Although RT can be summarized as a person characteristic by aggregating item-level information (see, e.g. OECD, 2020; Shin et. al., in press), , it is unclear how much dependency remains between individual item responses and corresponding RT variables. Therefore, M5 extends M4 by specifying additional parameters (π_j) to account for residual correlations among item responses and RT variables. In the equations, the difference between M4 and M5 is only shown in the item-level variance-covariance matrix while the latent regression part is kept the same in the following equation:

$$\begin{bmatrix} x_j^* \\ t_j^* \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} a_j\theta + bj \\ \beta_j - \gamma_j\tau \end{pmatrix}, \begin{pmatrix} 1 & \pi_j\alpha_j \\ \pi_j\alpha_j & \alpha_j^2 \end{pmatrix}\right], \\ \theta \mid \mathbf{z}, \tau \sim \mathcal{N}(\mathbf{\Gamma}'\mathbf{z} + \omega\tau, \sigma^2). \tag{14}$$

For all analyses, the official sampling weights were used, which are adjusted to represent the sampling design and non-response rates (Rust and Johnson, 1992). All five models were fit through WLSMV method using Mplus 8.0 (Muthén and Muthén, 2017) through the *MplusAutomation* package in R (Hallquist and Wiley, 2018). The method WLSMV involves diagonally weighted least-squares estimation with mean- and variance-adjusted chi-square statistics and standard errors. This method is more appropriate for analyzing categorical data such as item responses (Asparouhov and Muthen, 2007; Muthén et al., 1997), particularly when estimating correlations between categorical variables (RA) and continuous variables (RT) is of interest.

Results

To address our three research questions, we compared parameter estimates and model fit between nested models. First, we compared M1 and M2 to evaluate whether the within-item RA-RT conditional dependencies were significantly different from zero (RQ1). Next, a comparison between M3 and M4 showed the empirical consequences including RT in the IRT-LRMs to the measurement of ability (RQ2). Finally, the empirical consequences of ignoring conditional dependencies were addressed by comparing M4 and M5 in the IRT-LRM (RQ3).

RQ1. Presence of RA-RT conditional dependencies

Across all ten NAEP timing sections, the measurement model accounting for the within-item conditional dependencies (M2), fit significantly better than the measurement model that did not (M1; $\Delta\chi^2 s = 1935.29 - 14730.76, \Delta dfs = 15 - 21, ps < .0001$). To quantify the degree to which M2 fit better than M1, we compared the model fit statistics, Root Mean Squared Error of Approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980) and Comparative Fit Index (CFI; Bentler, (1990). The differences in RMSEA ranged from 0.001 to 0.008, and the differences in CFI ranged from − 0.139 to − 0.009, all favoring M2. As a guideline for testing measurement invariance, Chen (2007) recommended a change of ≤ − 0.01 in CFI combined with a change of ≥ 0.015 in RMSEA to indicate measurement non-invariance when the sample size is adequate (total $N > 300$). While the difference in CFI exceeded the threshold for nine out of ten NAEP timing sections, the RMSEA differences did not reach the threshold, meaning that the comparison did not meet the criteria for non-invariance. All PISA timing sections yielded similar results: significant test statistics across all 34 timing sections ($\Delta\chi^2 s = 136.38 - 673.13, \Delta dfs = 25 - 35, ps < .0001$). Yet, a comparison of other model fit statistics did not show substantial benefits of M2 over M1. Across 34 timing

Table 2 Residual variance estimates: NAEP 2017 data

Timing section	M3	M4	M5	M4 – M3	M5 – M4
T1	0.721 (0.009)	0.616 (0.010)	0.654 (0.009)	0.105	− 0.038
T2	0.719 (0.006)	0.609 (0.009)	0.615 (0.009)	0.110	− 0.006
T3	0.746 (0.009)	0.666 (0.010)	0.674 (0.010)	0.080	− 0.008
T4	0.737 (0.009)	0.572 (0.009)	0.574 (0.009)	0.165	− 0.002
T5	0.736 (0.009)	0.644 (0.009)	0.654 (0.009)	0.092	− 0.010
T6	0.725 (0.009)	0.620 (0.009)	0.630 (0.009)	0.105	− 0.010
T7	0.719 (0.009)	0.588 (0.010)	0.611 (0.009)	0.131	− 0.023
T8	0.733 (0.009)	0.630 (0.009)	0.633 (0.009)	0.103	− 0.003
T9	0.710 (0.009)	0.595 (0.009)	0.600 (0.009)	0.115	− 0.005
T10	0.702 (0.008)	0.618 (0.008)	0.624 (0.008)	0.084	− 0.006
Mean	0.725	0.616	0.627	0.109	− 0.011

M3 is the baseline model, including only contextual variables in the IRT-LRM.

M4 has all the parameters of M3, but also latent speed in the IRT-LRM.

M5 has all the parameters of M4, but also within-item parameters for all RA-RT conditional dependencies.

sections in PISA, differences in RMSEA ranged between 0.000 and 0.001, and the differences in CFI ranged between − 0.027 and − 0.002. Based on Chen's criteria, the threshold for CFI was exceeded in 16 of the 34 timing sections, but the change in RMSEA never reached the threshold.

RQ2. Inclusion of RT in the IRT-LRMs

Next, we evaluated the benefits of including RT variables in the IRT-LRMs. If incorporating the RT variables in the IRT-LRM improves measurement precision of ability (θ), this improvement can be quantified as the extent to which the residual variance in ability is reduced (M3 vs. M4) given that comparable constraints on model identification are used. Table 2 and Fig. 3 present the residual variances estimated in NAEP and PISA data. In both programs, inclusion of RT in the IRT-LRM resulted in a substantial increase in measurement precision of ability (i.e., reduced residual variance): about 11% in NAEP data and 17.6% in PISA data. This increased precision is in line with previous studies that reported improved reliability when RT was included in the measurement model (Bolsinova and Tijmstra, 2018; Reis Costa et al., 2021).

RQ3. Consequences of ignoring conditional dependencies

Given the presence of within-item conditional dependencies in NAEP and PISA, we examined impacts of conditional dependencies on the estimates in the IRT-LRM (*Assumption 5*). If the impacts of within-item RA-RT are substantial, regression coefficients estimated from M5 will be different from the ones estimated from M4. Therefore, we compared the regression parameter estimates between the IRT-LRM with (M5) and without (M4) parameters for the conditional dependencies: (a) the regression coefficient estimates for all the contextual variables excluding the regression coefficient for latent speed, (b) the residual variance estimates, and (c) the regression coefficient estimates for latent speed.

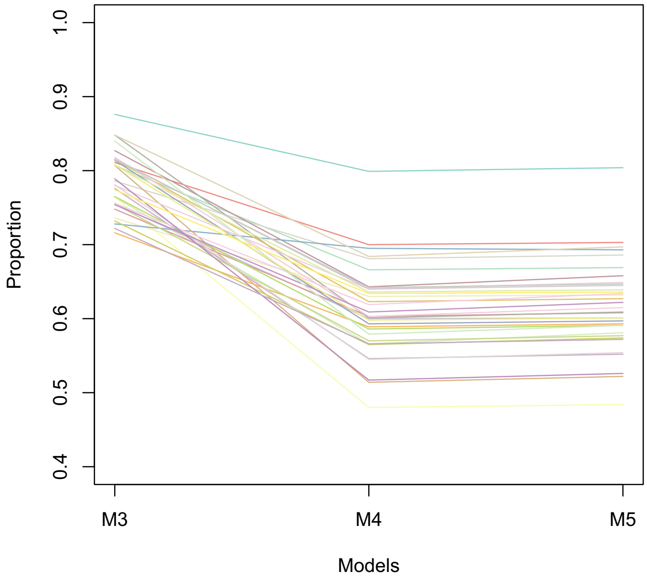

Fig. 3 Residual variance estimates in PISA data across IRT-LRMs (M3, M4, M5)

The regression coefficient estimates for the contextual variables were compared through root mean square deviation (RMSD) to evaluate whether these parameter estimates were different in M4 as a result of neglecting conditional dependencies. The RMSD was calculated as the square root of the mean of the squared differences between the estimates for M4 and M5. The RMSD values for the regression coefficients across all timing sections ranged from 0.000 to 0.001 in NAEP, and from 0.001 to 0.004 in PISA. Based on this analysis, it was apparent that the differences are negligible, and the same regression coefficient estimates were obtained regardless of whether the within-item RA-RT conditional dependencies were modeled.

In terms of the explained variance in latent ability, ignoring within-item conditional dependencies in M4 as compared to M5 resulted in only slightly lower residual variance estimates (about 1%) in both data sets (Table 2 and Fig. 3). The slightly lower residual variance estimates corresponded to the slightly larger-in-magnitude regression coefficient estimate for the latent speed (ω) in M4. Specifically, the regression coefficient estimates for the latent speed ranged between − 0.41 and − 0.29 in M4 in NAEP, and corresponding values for M5 ranged between − 0.30 and − 0.26. Similarly, in PISA, coefficient estimates ranged between − 0.55 and − 0.18 in M4, while the range was between − 0.54 and − 0.19 in M5. On average, ignoring within-item RA-RT conditional dependencies resulted in a slightly lower coefficient for τ of .015 on average, and the difference was negative (larger in magnitude) in every dataset. The negative estimates of the regression coefficients result from faster responses being associated with lower accuracy.

Concluding remarks

The evaluation of conditional independence assumptions underlying the IRT-LRM is important to ensure accurate estimation of group-level scores in LSAs. Such assumptions are more plausible if the contextual variables included in the latent regression are measured independently from the performance on the test (RA). However, conditional independence may be less plausible when it comes to incorporating process data derived from the cognitive assessment, such as response time, into the IRT-LRM. As response time and other types of process data are not measured independently from performance on the test, some degree of conditional dependence can be expected (Bolsinova et al., 2017).

Nevertheless, including the process data in the IRT-LRM is still vital to support secondary analyses involving relationships between process data and abilities with plausible values (Mislevy, 1991), in order to avoid estimation biases associated with violations of *congeniality* (Meng, 1994). In addition, including process data in the IRT-LRM can contribute to more precise estimation of latent ability (Bolsinova and Tijmstra, 2018; Reis Costa et al., 2021; Shin et al., in press). Therefore, we attempted to contribute to the field by focusing on the conditional dependencies in the context of LSAs. First, we evaluated RA and RT dependencies with NAEP and PISA datasets to examine the issue of whether the dependencies are also found in LSA data. Second, we evaluated the impact on the operational models used by LSAs (i.e., IRT-latent regression) when these models are fit to data that exhibit conditional dependencies. Through empirical analysis, we evaluated the benefits (i.e., increase in measurement precision) and the costs (i.e., potential estimation biases) of incorporating RT into the IRT-LRM, which has critical implications to the modelling of response time, and indirectly to other types of process data more generally, in LSAs. We conclude the paper with a summary of the main findings with suggestions for future studies.

First, we elaborated five types of conditional independence assumptions imposed in IRT-LRMs that jointly model RA and RT variables. In such models, two types of conditional independence assumptions are made that involve only a latent variable for RA (i.e., ability, θ). To jointly model latent variables for RA (ability, θ) and RT (speed, τ) based on van der Linden's (2007) hierarchical model, three assumptions were additionally required in the IRT-LRM. Most critical for the present paper is the assumption of conditional independence between RA and RT at the item-level given the latent variables in the measurement model and in the structural model.

To address the first research question, we evaluated within-item conditional dependencies in the measurement model. When parameters were added to the model to account for RA and RT conditional dependencies within each item, statistically significant within-item conditional dependencies exist in both NAEP and PISA, but their impact on the overall-model fit looks negligible. While a substantive explanation of conditional dependencies would be relevant, it is beyond the scope of this paper. Bolsinova et. al. (2017) and De Boeck and Jeon (2019) provide possible sources of observed RA-RT conditional dependencies, and show how explanatory models can help in investigating particular phenomena involved in the observed RA-RT conditional dependencies. For example, item types (e.g., response format), position effects (e.g., running out of time), the working speed of respondents (Fox and Marianti, 2016), attention variation or the change of problem-solving strategies during the test can be studied in the future as between-item dependencies. In that line of research, a closer look at item-level and person-level results will be worthwhile to understand the reasons behind the conditional dependencies. As Molenaar et. al. (2015) illustrated, modification indices can be useful for such purposes to identify items that show large residual correlations within or between items, given that all analyses considered in this paper can be understood as special cases of generalized linear latent variable models.

Next, we investigated the benefits of adding RT variables in the structural part of the IRT-LRM. Adding RT to IRT-LRM led to substantial improvement in measurement precision, approximately 11% in NAEP and 18% in PISA on average, over and above the key reporting variables. Here, the gain in measurement precision was quantified as the relative difference in residual variance estimates of θ after including RT in the structural model along with key reporting variables. Further studies can evaluate the impact on the reporting outcomes by taking into account imputation errors in generating the plausible values. Another research topic is related to the fit of the log-normal factor model for RT (van Sinharay and Rijn, 2020). More flexible approaches discussed in Entink, van Der Linden, and Fox (2009) and Glas and van der Linden (2010) could result in better fit to the distributions of the RT variables. However, if many additional parameters are needed to obtain good fit, this may complicate operational analysis of LSAs.

Finally, given the benefits of adding RT variables in the IRT-LRM, we quantified the cost of ignoring the conditional dependencies between RA and RT. Differences in regression coefficient estimates were small between IRT-LRMs that account for conditional dependencies and the corresponding models that did not. Ignoring within-item conditional dependencies resulted in slightly higher estimates of the regression coefficients for latent speed (about 0.015), corresponding to slightly lower residual variance estimates (about 1%) in both data sets. That is, the difference in residual variance estimates due to neglecting conditional dependencies were evidently smaller compared to the decrease in residual variance estimates by including RT to the IRT-LRM.

In summary, statistical evidence was found for RA-RT within-item conditional dependencies, consistent with previous research (e.g., Bolsinova, et. al. 2017), presenting a challenge for inclusion of RT in the operational models widely used in LSAs. However, speed was strongly correlated with ability, over and above the key reporting variables, suggesting that inclusion of RT in the IRT-LRM may be important to support secondary user analysis of RT to avoid congeniality-related violations (Meng, 1994). Furthermore, the observed reduction in the residual variance indicates that the inclusion of RT in IRT-LRMs improves the estimation of latent ability (e.g., Shin et al., in press). In contrast, only a relatively modest estimation difference was observed in the regression parameters from neglecting the within-item conditional dependencies in the IRT-LRM. Therefore, we conclude that the benefits of incorporating RT in the operational models for large-scale educational assessments may outweigh the costs.

A. List of contextual variables (z) in NAEP and PISA

In NAEP, LEP is a dichotomous variable that indicates Limited English Proficiency. IEP is a dichotomous variable that indicates an Individualized Education Plan. CENSREG indicates Census Region with categories of North East, Midwest, South, and West. DSEX is a dichotomous variable that indicates gender. SCHTYPE indicates school type, with categories of Public, Private, Catholic, Bureau of Indian Education school, and Department of Defense school. SLUNCH1 indicates eligibility for the National School Lunch Program, with categories of Eligible, Not Eligible, and Information not Available. SRACE10 indicates race/ethnicity, with categories of White, Black, Hispanic, Asian, American Indian or Alaskan Native, Native Hawaiian and Pacific Islander, and Mixed Race. PARED indicates student-reported highest educational attainment of either parent, with levels Some Highschool, Graduated Highschool, Post Higshool, Graduated College, and I don't know.

In PISA, HISCED is a variable that indicates the highest education level of parents, with categories of None and ISCED 1 combined, ISCED 2, ISCED 3B and ISCED 3A combined, and ISCED 5B, ISCED 5A, 6 combined due to the insufficient sample sizes. GENDER is a dichotomous variable (ST004D01T) that indicates gender. IMMIG represents immigration status, with categories of Native, First generation, and Second generation. NBOOKS represents the number of books (ST013Q01TA), with categories of 0-10 books, 11-25 books, 26-100 books, 101-200 books, 201-500 books, and more than 500 books.

All variables were contrast-coded, and the dichotomous contrast-coded variables were included in the regressions.

Acknowledgements
The authors would like to thank Emily Kerzabi for her editorial help.

Author contributions
HS analyzed and interpreted the PISA data and was a major contributor in writing the manuscript. PJ analyzed and interpreted the NAEP data and was a major contributor in writing the manuscript. PvR interpreted the results together and was a contributor in writing the manuscript. All authors read and approved the final manuscript.

Funding
Funding information is not applicable. This study (analysis, interpretation, writing the manuscript) was funded by Educational Testing Service through internal research allocation.

Availability data and materials
The NAEP dataset analyzed in the current study is available on request through NCES website (https://nces.ed.gov/pubsearch/licenses.asp). The PISA 2015 dataset analyzed in the current study is available to download through OECD website (https://www.oecd.org/pisa/data/2015database).

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 19 July 2021 Accepted: 2 June 2022
Published online: 22 June 2022

References

Asparouhov, T., & Muthen, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. proceedings of the 2007 jsm meeting in salt lake city, utah, section on statistics in epidemiology (2531–2535).

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238.

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics, 44*(6), 706–732.

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. https://doi.org/10.1111/bmsp.12104

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.

Bolsinova, M. , Tijmstra, J. , Molenaar, D. & De Boeck, P. (2017). Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. *Frontiers in psychology, 8*202.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504.

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology, 70*(2), 225–237.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102.

Entink, R. K., van Der Linden, W., & Fox, J.-P. (2009). A box-cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*(3), 621–640.

Ercikan, K. & Pellegrino, J W. (Eds). (2017). Validation of score meaning for the next generation of assessments: The use of response processes. Taylor & Francis.

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate behavioral research, 51*(4), 540–553.

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., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.

Greiff, S., Wstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Hallquist, M. N., & Wiley, J. F. (2018). Mplusautomation: an r package for facilitating large-scale latent variable analyses in m plus. *Structural equation modeling: a multidisciplinary journal, 25*(4), 621–638.

He, Q., von Davier, M., & Han, Z. (2018). Exploring process data in problem-solving items in computer-based large-scale assessments. In H. Jiao, R. W. Lissitz, & A. Van Wie (Eds.), *InData analytics and psychometrics*.NCInformation Age Publishing. (53–75)

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, G. Eugenio, K. Irwin, & Y. Kentaro (Eds.), *In TThe role of international large-scale assessments: perspectives from technology, economy, and educational research*. The NetherlandsSpringer. (1—11)

Lord, F. (1952). A theory of test scores. Psychometric monographs.

Lu, J., & Wang, C. (2020). A Response Time Process Model for Not-Reached and Omitted Items. *Journal of Educational Measurement, 57*(4), 584–620. https://doi.org/10.1111/jedm.12270

Mazzeo, J. & von Davier, M. (2008) .Review of the programme for international student assessment (pisa) test design: Recommendations for fostering stability in assessment results. Education Working Papers EDU/PISA/GB. *28*, 23–24.

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.

Meng, X-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*, 538–558.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Messick, S., Beaton, A., & Lord, F. (1983). *Naep reconsidered: a new design for a new era (naep report 83–1)*. Princeton, NJ: Educational Testing Service.

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The Relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing, 20*(3), 187–205. https://doi.org/10.1080/15305058.2019.1706529

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in naep. *Journal of Educational Statistics, 17*(2), 131–154.

Molenaar, D., & de Boeck, P. (2018). Response mixture modeling accounting for heterogeneity in item characteristics across response times. *Psychometrika, 83*(2), 279–297.

Molenaar, D. , Tuerlinckx, F. & VanderMaas, H.L.J. (2015).A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. Multivariate Behavioral Research 50, 56–74. https://doi.org/10.1080/00273171.2014.962684

Muthén, B. , du Toit, S. & Spisic, D. (1997).Robust interference using weighted least squares and quadratic estimating equations in the latent variable modeling with categorical and continuous outcomes. Unpublished manuscript, University of California

Muthén, B., & Muthén, L. (2017). *InMplus Mplus*. Chapman and Hall/CRC.

NCES. (2018a). In2017 NAEP Mathematics Report Card(Tech. Rep.). https://www.nationsreportcard.gov/math_2017/

NCES. (2018b). In2017 NAEP Reading Report Card(Tech. Rep.). https://www.nationsreportcard.gov/reading_2017

OECD. (2017). PISA 2015 Technical Report. Paris, France: OECD Publishing.

OECD.2020. *PISA (2018) technical report (annex H: new procedures for PISA 2018 population modelling)*. FranceOECD Publishing.

Partchev, I. & De Boeck, P. (2012) .Can fast and slow intelligence be differentiated? *Intelligence,40*(1), 23–32.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of educational and behavioral Statistics, 24*(2), 146–178.

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika, 84*(3), 892–920. https://doi.org/10.1007/s11336-019-09669-2

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340.

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.

Reis Costa, D., Bolsinova, M., Tijmstra, J. Andersson, B. (2021). Improving the precision of ability estimates using time-on-task variables: insights from the PISA 2012 computer-based assessment of mathematics. *Frontiers in Psychology,* https://doi.org/10.3389/fpsyg.2021.579128

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association, 91*(434), 473–489.

Rust, K. F., & Johnson, E. G. (1992). Chapter 2: Sampling and weighting in the national assessment. *Journal of Educational Statistics, 17*(2), 111–129.

Shin, H. J., von Davier, M., & Yamamoto, K. (in press). Incorporating timing data into the PISA population modeling. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative Computer-based International Large-Scale Assessments - Foundations, Methodologies and Quality Assurance Procedures*. Springer.

van Sinharay, S., & Rijn, P. W. (2020). Assessing fit of the lognormal model for response times. *Journal of Educational and Behavioral Statistics, 45*(5), 534–568.

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate behavioral research, 25*(2), 173–180.

Teig, N., Scherer, R., & Kjærnsli, M. (2020). Identifying patterns of students' performance on simulated inquiry tasks using pisa 2015 log-file data. *Journal of Research in Science Teaching, 57*(9), 1400–1429.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*(S1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425–453. https://doi.org/10.1080/00273171.2019.1643699

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z

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 Rijn, P. W., & Ali, U. S. (2018). A generalized speed-accuracy response model for dichotomous items. *Psychometrika, 83*(1), 109–131.

van Rijn, P. W., & Ali, U. S. (2018). Sarm: A computer program for estimating speed-accuracy response models for dichotomous items. *ETS Research Report Series, 2018*(1), 1–18.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series, 2*(1), 9–36.

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.

von Davier, M. Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis, 155–174.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). 32 the statistical procedures used in national assessment of educational progress: recent developments and future directions. *Handbook of statistics, 26*, 1039–1055.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477.

Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling, 58*(4), 671–701.

Yamamoto, K., & Lennon, M. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education: An International Perspective, 26*(2), 196–212. https://doi.org/10.1108/QAE-07-2017-0038

Publisher's Note

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