JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12391
Journal of Educational Measurement
Fall 2024, Vol. 61, No. 3, pp. 378–407

Modeling the Intraindividual Relation of Ability and Speed within a Test

Augustin Mutak
Freie Universität Berlin, Berlin, Germany
Robert Krause
University of Kentucky, Lexington, United States
Esther Ulitzsch
University of Oslo, Oslo, Norway
Sören Much and Jochen Ranger
Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
Steffi Pohl
Freie Universität Berlin, Berlin, Germany

Understanding the intraindividual relation between an individual's speed and ability in testing scenarios is essential to assure a fair assessment. Different approaches exist for estimating this relationship, that either rely on specific study designs or on specific assumptions. This paper aims to add to the toolbox of approaches for estimating this relationship. We propose the intraindividual speed-ability-relation (ISAR) model, which relies on nonstationarity of speed and ability over the course of the test. The ISAR model explicitly models intraindividual change in ability and speed within a test and assesses the intraindividual relation of speed and ability by evaluating the relationship of both latent change variables. Model estimation is good, when there are interindividual differences in speed and ability changes in the data. In empirical data from PISA, we found that the intraindividual relationship between speed and ability is not universally negative for all individuals and varies across different competence domains and countries. We discuss possible explanations for this relationship.

Introduction

With computerized testing becoming more and more common, there has been a trend of leveraging additional data, such as response times, in addition to the examinee's responses. Response time data may provide in-depth insights into the examinees' behavior during the test. For instance, they can be employed as measurement indicators of examinee's speed, modeled alongside customary ability estimates. Response time data can be used to better understand multiple aspects of examinee behavior, such as test-taking strategies, motivation, and cheating (Kyllonen & Zu, 2016). In this paper, we model response time alongside item response information to investigate the intraindividual relationship between ability and speed in psychometric assessments.

One of the most widely used models for joint psychometric analysis of responses and response times is the speed-accuracy (SA) model by van der Linden (2007). The model describes two levels. On the lower level, it is composed of measurement models for latent ability and latent speed. On the second level, the model relates these two latent variables, which allows researchers to explore the interindividual relationship between speed and ability. The model has served as a basis for many extensions. It has been extended to include multiple dimensions in the ability measurement model (Man et al., 2019), allowing for the exploration of compensatory effects of different ability dimensions along with the response times. Other extensions include models which aim to increase the precision of ability estimates (Bolsinova & Tijmstra, 2018; Molenaar, Oberski, Vermunt, & De Boeck, 2016; Molenaar, Tuerlinckx, & van der Maas, 2015; Molenaar, Bolsinova, & Vermunt, 2018; van der Linden, 2008; van der Linden & Guo, 2008), account for missing values (Frey, Spoden, Goldhammer, & Wenzel, 2018; Liu & Wang, 2020; Liu, Wang, & Shi, 2021; Pohl, Ulitzsch, & von Davier, 2019; Ulitzsch, von Davier, & Pohl, 2020c, 2020b), model guessing behavior (Guo et al., 2016; Schnipke & Scrams, 1997; Wang & Xu, 2015), or to model engaged and disengaged test-taking behavior (Ulitzsch, von Davier, & Pohl, 2020a).

One extension of van der Linden's (2007) model which is particularly noteworthy for the context of this paper is the one by Fox and Marianti (2016). In this model extension, the authors relax the assumption of stationarity of speed and model intraindividual change in speed over the course of the test using latent growth modeling. Peng, Cai, Wang, Luo, & Tu (2022) build upon the model of Fox and Marianti (2016) and applied it to cognitive diagnostic modeling. They sorted participants into multiple groups with regards to the measured attributes and examine how each of these groups behaves when it comes to their speed fluctuations within the test. Both, Fox and Marianti (2016) as well as Peng et al. (2022), do, however, assume that even though speed may change within a person, the ability level stays the same. This is a rather strong assumption that may not hold in applications. In fact, lately different research has dealt with the relation of ability and speed on the intraindividual level. This is not only relevant to evaluate model assumptions (e.g., of stationarity or conditional independence), but is also important as differences in speed may also impact ability level and is, thus, relevant for a fair comparison of persons (e.g., Pohl, Ulitzsch, & von Davier, 2021).

Intraindividual Relationship between Ability and Speed

In psychometrics, it has been acknowledged that exhibited ability may depend on the speed with which a person works on the test (e.g., Goldhammer, 2015; van der Linden, 2007). As such a person may exhibit different ability levels, depending on the speed they chose. Ranger, Kuhn, & Pohl (2021) define the hypothetical maximum level of ability which an examinee could obtain as *target ability*. This is not necessarily observed in a specific assessment. Instead, we usually assess *effective ability*, that is, the ability exhibited at the chosen, that is the effective speed of the test taker (Goldhammer, 2015). While each person has one level of target ability, they have multiple values on effective ability, depending on the effective speed level. If

persons do not change in speed across the test, we usually observe only one effective speed and effective ability level per person.

Psychometric research has emerged that investigates the intraindividual relationship between ability and speed in various different ways (e.g., Alfers, Gittler, Ulitzsch, & Pohl, 2021; Kang, De Boeck, & Ratcliff, 2022; Ranger, Kuhn, & Pohl, 2021). Research aiming to investigate this relationship in psychometric assessments is challenged by difficulties in identifying this relation on the intraindividual level. Estimating the intraindividual relationship between ability and speed requires one to have several values of both speed and ability for each person. Different strategies have been proposed to tackle this. These either rely on (a) using experimental designs, (b) using external proficiency measures, or (c) relying on nonstationarity of ability and speed within the test.

Experimental designs. To investigate the intraindividual relationship of ability and speed with an experimental design, researchers usually manipulate the time limits for the test (Alfers et al., 2021; Nietfeld & Bosma, 2003) or each item response (Goldhammer, 2015). The advantage of experimental approaches is that they assure a variation of speed within persons and have good internal validity. Experimental variations may, however, be confounded by order or position effects. Manipulating an examinee's working speed requires either a within-subject design, which might interfere with order and position effects or a between-subject design, which relies on comparability of persons across groups. In practice, it may also often be unfeasible to implement experimental conditions, which makes this approach less applicable to many studies which are assessments.

External proficiency measures. Another method to infer the intraindividual relation of ability and speed is to make use of external proficiency scores. For example, Ranger et al. (2021) used data from the Amsterdam Chess Test (van der Maas & Wagenmakers, 2005) to obtain participants' effective speed and effective ability. They used the Elo score as an external measure of proficiency and grouped persons with a similar Elo score. Assuming that persons with a similar Elo score have a similar intraindividual relation of ability and speed, they thus, were able to estimate the intraindividual relationship of speed and ability. This approach is easier to apply than an experimental manipulation since it does not require a controlled environment. However, to use this approach, one needs an external measure of proficiency that is not affected by any choice of speed level. Such a measure is rarely available in practice. Furthermore, the assumption that examinees with similar proficiency show the same relationship between ability and speed, is likely to be violated in practice.

Nonstationarity. Research has shown that speed of a person is not necessarily stable within a test, but that a person may change work pace (e.g., Fox & Marianti, 2016). Due to the speed-ability trade-off (SAT), a change in effective speed may also impact test performance and as such effective ability. Some approaches aimed at investigating the intraindividual relationship between ability and speed make use of nonstationarity of speed and ability across the test. Nonstationarity allows for observing different levels of effective speed and effective ability within the same person.

For example, Domingue et al. (2022) used an approach that relied on residuals. They first modeled responses assuming stationarity of ability. They then regressed the residuals of the responses on the response times to evaluate in which way deviations from stationarity are impacted by response times. Similar as Domingue et al. (2022), Guo, Luo, and Yu (2020) applied an approach in which a single latent speed and ability variable are modeled. In order to infer the intraindividual relationship of ability and speed, different from Domingue et al. (2022), they regress the responses on the standardized residual response times. Also, Meng, Tao, and Chang (2015) relied on nonstationarity of ability and speed and modeled conditional dependencies by extending the model of van der Linden (2007). In contrast to the other two models, they allowed for different strength of the intraindividual residual correlation across persons. Also in the diffusion IRT (Item Response Theory) model with random variability coefficients (DIRT-RV, Kang et al., 2022), it is allowed that the size and even the form of the relationship differs between persons. In this model, however, it is assumed that the form and size of the relationship depends on the ability level. Specifically, it is assumed that the relation of accuracy and response time is opposite for high response probability than for low response probability. This assumption has shown to be less plausible in psychometric test data (Krause et al., 2022).

Approaches relying on nonstationarity have the advantage of being applicable to data most often found in practice, as they neither need experimental manipulation nor external measures. The only requirement is that changes in speed and ability do occur within persons across the test. As such, they are very promising for wide use. While the existing approaches provide already great tools for investigating the intraindividual relationship, they mainly do not directly model the relation of ability and speed, but only the relation of the residuals. This only indirectly allows to infer to the speed-ability relation.

Results of Investigating the Intraindividual Relationship of Speed and Ability in Psychometric Assessments

The results of studies investigating the intraindividual relationship of speed and ability in psychometric data show mixed results. There are a number of studies that suggest that the intraindividual relationship between both variables is negative. Guo et al. (2020) found support for the negative within-person relationship of effective speed and effective ability for almost all items but those of middle difficulty. The study by Nietfeld and Bosma (2003) shows that, while most of the participants display patterns of a negative relationship between ability and speed, there is also a considerable number of individuals who do not fit into the pattern (their responses are fast and accurate or slow and inaccurate). Domingue et al. (2022) analyzed 29 different data sets with different kinds of tests and found inconsistent results on the estimated within-person relation of speed and ability. A negative relationship was found for some tests, but not for others. Noticeably, positive relationships were mainly found in data sets consisting of participants of higher age. For certain tasks, a curvilinear dependency was found—ability first rises, and then declines as speed increases. Kang et al. (2022) also found that, in general, there is a curvilinear within-person relationship of effective speed and effective ability on mental rotation tasks.

The curvilinear dependency was of the same shape—ability first increases, and then decreases with the increase in speed. Similar results were obtained by Ranger et al. (2021), who found that for some people the relationship between effective speed and effective ability is curvilinear. However, they also found that in other groups the relationship of effective speed and effective ability was positive until it reaches a certain plateau. They also find a dependency on overall ability level; for respondents with above-average ability, the relationship between effective speed and effective ability is positive, although above a certain level of speed, this relationship becomes weaker or nonexistent.

Explanations for the Intraindividual Relationship of Ability and Speed

There are different explanations of reasons for the intraindividual relation of ability and speed within a test (e.g., Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017; Goldhammer, 2015; Pohl et al., 2021; Ranger et al., 2021; van Breukelen, 2005).

The SAT. One of the most prevalent explanations for the intraindividual relationship of ability and speed is the SAT. The SAT refers to the widely observed within-person decrease in task performance which appears as a consequence of an increase in the speed of performing the task. It can be visualized with a SAT curve, which shows the relationship between speed and ability as a monotonically decreasing function. The SAT curve is thought to be asymptotic (Goldhammer, 2015), that is, effective ability levels off after some time.

Because persons can show different levels of speed, and thus, of ability, van der Linden (2007) introduced the terms of *effective* ability and *effective* speed. As in psychometric testing, different examinees usually do not use the same speed level, but differ in their choice, fairness of comparisons of ability levels across examinees are threatened (Goldhammer, 2015; Pohl et al., 2021). As such, investigation of SAT is an important topic when it comes to fairness in assessments.

Changes in concentration. A positive relationship between speed and ability could be a result of changes in concentration of the participant across the test (Ranger et al., 2021). When the concentration of an individual decreases, they experience an increase in task-irrelevant cognition. As a result, the participant spends more time on an item, given that a part of the time is used up for task-irrelevant cognition. Simultaneously, not focusing well enough on the task deteriorates the participant's performance. If a participant's concentration changes during the test, both speed and ability may change in the same direction, thus resulting in a positive relationship between the two (Bolsinova et al., 2017).

Changes in effort. The effort an examinee invests into solving the items can impact both, ability and speed. If effort changes across the course of the test, this can have an impact on the intraindividual relationship of their speed and ability. Investing more effort can lead to longer response times, given that a larger amount of information is processed. At the very least, extremely fast response times are thought to reflect rapid guessing (Wise & Kong, 2005), and are thus indicative of nonmotivated test-taking behavior. On the other hand, whether the increase in effort will lead to higher accuracy or not depends on the capability of the examinee (limited by the

maximal ability level of the examinee; Ranger et al., 2021). In examinees capable enough to solve an item, investing more effort should also lead to better accuracy, while in examinees not capable enough to solve an item correctly, the effect should be opposite. Likewise, a faulty solution process might lead to lower speed and lower accuracy (van Breukelen, 2005). Thus, the direction of the relationship of speed and accuracy also depends on the person's capability. For a capable examinee, a change in effort results in a negative relationship, while for a less capable examinees a change in effort results in a positive relationship.

Practice effects. Practice effects are yet another factor which could influence the intraindividual relationship of speed and ability. Practice effects refer to the changes in the solving process which take place after the examinee got acquainted with the nature of the test material. It has been shown that practice leads to an increase in accuracy (Scharfen, Peters, & Holling, 2018), although the effect reaches a plateau after a certain point. Practice was also shown to increase examinees' speed (Scharfen, Blum, & Holling, 2018). Thus, an increase practice effects during the course of the test would increase both speed and accuracy, and, thus, result in a positive intraindividual relationship between the two.

Research Objectives

The current study aims to add another approach to the toolbox of approaches for modeling the intraindividual relationship of speed and ability. Our model does not aim to be always superior to all other approaches but rather to provide another tool with different strengths and limitations than the existing ones. By adding to the toolbox, a researcher may choose which of the approaches best fits their aim and which of the assumptions is most plausible in a given situation.

We draw on modeling approaches that rely on nonstationarity (such as Domingue et al., 2022; Guo et al., 2020), as these are most widely applicable in many assessment settings, that is, they do neither require experimental manipulations (as for example, in Nietfeld & Bosma, 2003; Alfers et al., 2021; or Goldhammer, 2015), nor external measures of proficiency (as for example in Ranger et al., 2021).

Building upon modeling ideas of Fox and Marianti (2016) and Meng et al. (2015), we also aim to explicitly model change in speed, and thus, enable to model and explain this change in further analyses. In contrast to Fox and Marianti (2016), Meng et al. (2015), and similar models such as Peng et al. (2022), we do not assume that ability is stationary but at the same time allow for ability to also change across the test and for change in ability also to depend on change in speed.

While similarly as Domingue et al. (2022), Guo et al. (2020), Meng et al. (2015), and Kang et al. (2022), we rely on nonstationarity, different from them, we explicitly model the relationship of ability and speed (and not of residuals). This also allows for including further explaining variables for the change in ability and speed.

In the following, we will present (1) the proposed model, (2) a simulation study evaluating the performance of model estimation, and (3) an empirical application illustrating the use of the model.

Model

We are proposing a model that explicitly models changes in effective speed and effective ability and allows for investigating their within-person relation. For this, we rely on nonstationarity of effective speed and effective ability throughout the test. The proposed model is based on the model of Fox and Marianti (2016) which allows the growth in latent speed but assumes stationarity of ability.

Fox and Marianti (2016) based their approach on the hierarchical model of van der Linden (2007). The first level includes measurement models for ability and for speed. Ability is modeled via an IRT model, and a lognormal distribution is assumed for the response times. On the higher level, the model features a multivariate normal distribution of latent ability and speed.

Fox and Marianti (2016) extended this model by adding a growth model for the latent speed parameter. Specifically, they demonstrated the usage of linear and linear-quadratic growth terms. This allows researchers to look at person-level speed trajectories and to explore how speed trajectories are related to the person's ability level.

We extend the model of Fox and Marianti (2016) to also allow for changes in effective ability and change in effective ability and effective speed to be related within persons.

Model Specification

The intraindividual speed-ability-relationship (ISAR) model is specified as follows. On the lower level, the model consists of the measurement models for the responses and response times. The measurement model for the responses is given by:

$$\mathbb{P}(Y_{pi} = 1) = \frac{\exp(\theta_{0,p} + \theta_{1,p} X_{pi} - b_i)}{1 + \exp(\theta_{0,p} + \theta_{1,p} X_{pi} - b_i)}, \tag{1}$$

with

$$X_{pi} = \frac{l_{pi} - 1}{K}. \tag{2}$$

Y_{pi} denotes the response of examinee p to item i, $\mathbb{P}(Y_{pi} = 1)$ is the probability for a correct response and b_i is the difficulty parameter of item i. X_{pi} is the timescale variable which represents the relative position of item i in the test as encountered by examinee p. l_{pi} denotes the absolute position of item i as encountered by person p in the test starting with $l_{pi} = 1$ for i being the first item in the test for person p. K is the total number of items administered. In the case where all examinees encounter the items in the same order, l_{pi} and X_{pi} are reduced to l_i and X_i, respectively. The scale of X starts from 0 and its theoretical upper bound is 1, which it reaches when $K = \infty$ (see also Fox & Marianti, 2016). Due to this timescale, $\theta_{0,p}$ represents the effective ability of person p at the beginning of the test (initial ability) and $\theta_{1,p}$ represents the rate of change in effective ability of person p from the beginning of the test to the hypothetical $K + 1$th item (change in ability).

Following (Fox & Marianti, 2016), the measurement model for the response times is given by:

$$\ln(T_{pi}) \sim \mathcal{N}(\beta_i - (\tau_{0,p} + \tau_{1,p}X_{pi}), \sigma^2{}_{T_i}), \tag{3}$$

where T_{pi} is the response time of person p to item i, β_i is the time intensity parameter of item i, $\sigma^2{}_{T_i}$ is the residual variance of the log response times to item i, $\tau_{0,p}$ is the effective speed of person p at the beginning of the test (initial speed), and $\tau_{1,p}$ is the rate of change in effective speed of person p from the beginning of the test to the hypothetical $K + 1$th item (change in speed).

On the higher level, the latent speed and latent ability parameters are related to each other via a multivariate normal distribution:

$$\begin{pmatrix} \theta_0 \\ \theta_1 \\ \tau_0 \\ \tau_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\theta_0} \\ \mu_{\theta_1} \\ \mu_{\tau_0} \\ \mu_{\tau_1} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\theta_0} & \rho_{\theta_0\theta_1} & \rho_{\theta_0\tau_0} & \rho_{\theta_0\tau_1} \\ \rho_{\theta_0\theta_1} & \sigma^2{}_{\theta_1} & \rho_{\theta_1\tau_0} & \rho_{\theta_1\tau_1} \\ \rho_{\theta_0\tau_0} & \rho_{\theta_1\tau_0} & \sigma^2{}_{\tau_0} & \rho_{\tau_0\tau_1} \\ \rho_{\theta_0\tau_1} & \rho_{\theta_1\tau_1} & \rho_{\tau_0\tau_1} & \sigma^2{}_{\tau_1} \end{pmatrix} \right), \tag{4}$$

where the μ values are the means of the person parameters, σ^2 are their variances, and ρ are the correlations between them.

Similar as in Molenaar et al. (2015, 2016), we model item parameters (b_i and β_i) as fixed effects. The person parameters θ_0, θ_1, τ_0, and τ_1 are assumed to follow a multivariate normal distribution:

A path diagram of the ISAR model is depicted in Figure 1.

The crucial features of the ISAR model are the correlation $\rho_{\theta_1\tau_1}$, reflecting the relationship of the change in speed and the change in ability, and the individual lines depicting the relationship of speed and accuracy within every person.

To identify the model, we introduce two scaling constraints. First, we constrain the error variance of the log response times, $\sigma^2{}_{T_i}$, to be the same across all items ($\sigma^2{}_T$). Second, we fix all the person parameter means (μ_{θ_0}, μ_{θ_1}, μ_{τ_0}, and μ_{τ_1}) to 0. Note, that by posing this constraint, we cannot evaluate whether on average there is an increase or decrease in effective ability or effective speed, but we can only evaluate whether effective ability and effective speed change differently across persons. If all items are presented to all examinees in the same order, we cannot disentangle average change in effective ability and effective speed from differences in item difficulty or item time intensity, respectively. It is possible to disentangle average change in latent variables from item difficulty and item time intensity, when items are presented in random order to each person. In such settings, identification can be achieved by fixing the item difficulty and time intensity of one item to be equal across item positions, instead of fixing μ_{θ_1} and μ_{τ_1} to 0.

The proposed approach poses several requirements for its implementation and identification and makes several assumptions. First, as the timescale in the model is operationalized by item position, the model is only applicable to data in which persons approach the test in a linear way, without revising items. The estimation of the model as well as the investigation of the nonstationarity requires that persons indeed change their effective speed and effective ability throughout the test. In data in which this is not expected, the model may not be estimable. This is similar to

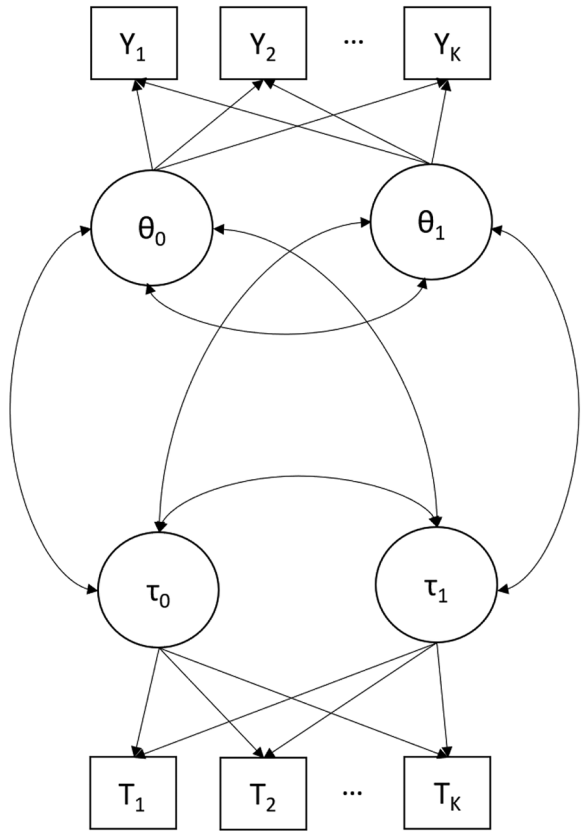

Figure 1. Path diagram of the ISAR model. Y and T represent the responses and response times, respectively, while the subscript $1, 2, \ldots, K$ denotes the ith item.

the requirements of the other approaches relying on nonstationarity. Similar as the model of van der Linden (2007), the model assumes (a) that item responses follow an IRT model and response times follow a lognormal model and (b) conditional independence of item responses and response times given the latent person variables. Although we do not assume stationarity of ability and speed, we still make assumptions on the form of the change, here being a linear one.

While the ISAR model cannot account for possible nonlinear relationships of speed and ability which might be present, it may offer some insight into the nature of the examinee's response process. For example, a positive relationship between speed and ability of an examinee would strongly suggest that the SAT did not play a major role when that examinee was giving their response. Likewise, a negative relationship between speed and ability of an examinee could potentially rule out any major changes in their concentration during the test or any practice effects.

Prior Distributions

Relying on Bayesian estimation, we specify the prior distributions for all parameters. For the item difficulty parameters b and the item time intensity parameters β we use weakly informative normal distribution priors with a mean of 0 and a standard deviation of 2 (Gelman et al., 2008). For the error variance of the log response times, a weakly informative half-Cauchy prior is chosen with a location of 0 and a scale of 2.5, which is a commonly used prior distribution for variances (Gelman, 2006). For the variance-covariance matrix Σ of person parameters, we use a separation strategy to avoid the problem of dependencies between the variances and the correlations (Alvarez, Niemi, & Simpson, 2014). Again, we use a half-Cauchy distribution with a location of 0 and a scale of 2.5 as a prior for the variances of the person parameters: For the correlation matrix of person parameters Ω, we choose an uninformative LKJ distribution (Lewandowski, Kurowicka, & Joe, 2009), specifically tailored for correlation matrices: $\Omega \sim LKJ(1)$. The variance-covariance matrix Σ may be calculated as

$$\Sigma = \text{diag}(\sigma_P)\Omega\ \text{diag}(\sigma_P), \tag{5}$$

where σ_P is the vector of variances of person parameters (${\sigma^2}_{\theta_0}$, ${\sigma^2}_{\theta_1}$, ${\sigma^2}_{\tau_0}$ and ${\sigma^2}_{\tau_1}$) and $\text{diag}(\sigma_P)$ is a diagonal matrix with the values of σ_P on the diagonal.

Estimation

In our implementation, we employed a Bayesian MCMC (Markov chain Monte Carlo) sampler in Stan (Stan Development Team, 2021), which utilizes Hamiltonian Monte Carlo techniques with a No U-Turn Sampler. We provide the Stan code in the Appendix.

Reparameterizations

Due to the number of correlated dimensions, estimation is more challenging. In order to enhance estimation, we reparameterized the model (Stan Development Team, 2021). First, we employ the Cholesky decomposition (Benoit, 1924) for easier estimation of the correlation matrix. The correlation matrix Ω can be written as a product of a lower triangular matrix L and its transpose L^T as follows:

$$\Omega = \underbrace{\begin{pmatrix} L_{11} & & & \\ L_{21} & L_{22} & & \\ L_{31} & L_{32} & L_{33} & \\ L_{41} & L_{42} & L_{43} & L_{44} \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} L_{11} & L_{21} & L_{31} & L_{41} \\ & L_{22} & L_{32} & L_{42} \\ & & L_{33} & L_{43} \\ & & & L_{44} \end{pmatrix}}_{L^T}, \tag{6}$$

where all the L_{cd} values (c denoting the row and d denoting the column in the matrix L) are called Cholesky factors. The Cholesky decomposition shows excellent numerical stability (Higham, 2009) and is thus more efficient than the usual parameterization which directly includes a correlation matrix (Stan Development Team, 2021).

Second, the specification of the distribution of person parameters was modified so that the parameterization is noncentered. To achieve this, a vector of intermediate

variables ${\theta_0}^*$, ${\theta_1}^*$, ${\tau_0}^*$, and ${\tau_1}^*$ was created. Each of the intermediate variables ${\theta_0}^*$, ${\theta_1}^*$, ${\tau_0}^*$, and ${\tau_1}^*$ was assumed to follow a standardized normal distribution with the mean of 0 and a standard deviation of 1. The original vector of person parameters was then decomposed as follows:

$$\begin{pmatrix}\theta_0\\ \theta_1\\ \tau_0\\ \tau_1\end{pmatrix} = \begin{pmatrix}\mu_{\theta_0}\\ \mu_{\theta_1}\\ \mu_{\tau_0}\\ \mu_{\tau_1}\end{pmatrix} + \begin{pmatrix}\sigma^2_{\theta_0}\\ {\sigma^2}_{\theta_1}\\ {\sigma^2}_{\tau_0}\\ {\sigma^2}_{\tau_1}\end{pmatrix} \left(L \begin{pmatrix}\theta_0^*\\ {\theta_1}^*\\ {\tau_0}^*\\ {\tau_1}^*\end{pmatrix}\right). \tag{7}$$

Thus, direct sampling from the multivariate normal distribution is avoided. By excluding explicit hierarchical correlations shown in Equation 4 from the sampling process and instead recalculating their values based on intermediate parameters, noncentered parameterization removes the dependence of the lower-order model parameters on the higher-order model parameters, instead making them both dependent just on the data during the process of sampling (see Papaspiliopoulos et al. (2007) for a detailed overview of the noncentered parameterization techniques and Neal (2003) for the mathematical background of the issues which may arise in estimating hierarchical models). Noncentered parameterization was shown to outperform centered parameterization and overcoming posterior pathologies (Betancourt & Girolami, 2015) and is a recommended choice for hierarchical models in Stan (Stan Development Team, 2021).

Third, the Cauchy prior for the variances was reparameterized by creating an intermediate variable. We here show this reparameterization on the example of the error variance of log response times. The original parameter, ${\sigma^2}_T$, is obtained as a derived parameter as

$${\sigma^2}_T = \gamma \tan({\sigma^{*2}}_T), \tag{8}$$

where γ is the scale parameter from the Cauchy prior (in our case, 2.5) and ${\sigma^{*2}}_T$ is the intermediate variable. The intermediate variable ${\sigma^{*2}}_T$ is sampled from a uniform distribution: ${\sigma^{*2}}_T \sim \mathcal{U}_{0,\frac{\pi}{2}}$. The variances of the person parameters are reparameterized in the same way. An intermediate variable is introduced for each of the four person parameters, and the original variable is computed as γ (2.5) times the tangent of the intermediate variable. All of the intermediate variables follow a uniform distribution bounded between 0 and $\frac{\pi}{2}$. These reparameterizations rely on using the tangent function to construct the cumulative distribution function of the Cauchy distribution, but avoid sampling from the heavy-tailed Cauchy distribution itself, which Hamiltonian Monte Carlo algorithms, as implemented in Stan, have trouble sampling from (Stan Development Team, 2021).

As all elements of the decomposition of the covariance matrix Σ (see Equation 5) are reparameterized, $\text{diag}(\sigma_P)$ may be calculated as $\text{diag}(\gamma \tan(\sigma_P{}^*))$, where $\sigma_P{}^*$ is a vector of intermediate variables used to reparameterize the Cauchy distribution of person parameter variances (${\sigma^2_{\theta_0}}^*$, ${\sigma^2_{\theta_1}}^*$, ${\sigma^2_{\tau_0}}^*$, and ${\sigma^2_{\tau_1}}^*$). The correlation matrix Ω is calculated as $L * L^T$, as shown in Equation 6.

Reparameterizing our model results in more regular shapes of trace plots and less bias in the estimated parameters than without these reparameterizations. They also

noticeably reduce the estimation time. However, we also found that a sufficiently large number of iterations was needed to obtain large enough effective sample sizes (ESS).

Simulation Study

In the simulation study, we examined under which conditions the parameters of the ISAR model can be accurately recovered in estimation. We specifically aim to investigate the minimum requirements and, thus, the boundary conditions for estimating the model.

Data Generation

Item responses and response times were generated according to Equations 1 and 3, respectively. Item difficulty and item time intensity parameters were simulated to be uncorrelated. The values were chosen to cover the most common range of the latent score scale. Item difficulty parameters were generated as $b_i = -1.5 + (i-1) * 3/(n_i - 1)$ with n_i giving the number of items. This results in difficulty parameters ranging from -1.5 to 1.5 in equidistant steps. This was similarly done for time intensity parameters β_i with the same values, that is, $\beta_i = -1.5 + (i-1) * 3/(n_i - 1)$. According to the results of Fox and Marianti (2016), the residual variance of response times ${\sigma^2}_T$ was set to .3. Person parameters were drawn from a multivariate normal distribution according to Equation 4 with the means of all four person variables set to zero. According to results from empirical studies (Fox & Marianti, 2016; Ulitzsch et al., 2020c, 2020b), we set the variance of the initial ability (${\sigma^2}_{\theta_0}$) to 1.15 and of the initial speed (${\sigma^2}_{\tau_0}$) to .3.

We varied four factors in the simulation study. These were the size of the variances of person parameters (small; large), the strength of the correlations of person parameters (weak; strong), the sample size (1,500; 5,000; 8,500), and the number of items/test length (9; 25). In the small variance condition, the variances of the latent variable slopes were 1/6 of the variances of their respective intercepts (i.e., ${\sigma^2}_{\theta_1} = .2$ and ${\sigma^2}_{\tau_1} = .05$), whereas in the large variance condition they were equal to the variance of the latent intercept variables (${\sigma^2}_{\theta_0} = 1.15$, ${\sigma^2}_{\tau_0} = .3$). To generate correlation matrices for the two correlation levels, we made use of the LKJ distribution (Lewandowski et al., 2009). We randomly drew values from the LKJ distribution to generate a correlation matrix Ω. For the weak condition, we set the η parameter of the LKJ distribution (which determines the strengths of correlation) to 3, and for the strong condition, we set it to .3. This resulted in the correlation matrices shown in Table 1.

The above setup yielded a total of $2 \times 2 \times 3 \times 2 = 24$ conditions. For each of the conditions, we generated 50 data sets. In total, there were 1,200 generated data sets. Model estimation was performed on Freie Universität Berlin's high-performance cluster *Curta* (Bennett, Melchers, & Proppe, 2020) in Stan (Stan Development Team, 2021). We have estimated the model in 10 chains with 10,000 iterations each, of which 4,000 are warm-ups.[1]

Table 1
Correlations of Person Parameters Used in the Simulation Study in the Weak and Strong Condition

	Weak				Strong		
	θ_0	θ_1	τ_0		θ_0	θ_1	τ_0
θ_1	−.13			θ_1	−.67		
τ_0	.17	−.09		τ_0	.27	−.17	
τ_1	.16	−.14	−.15	τ_1	.59	−.66	−.54

Evaluation Criteria

To evaluate model performance, we examined several criteria, both on model as well as on parameter level. To assess convergence of the estimation, we examined the $\widehat{R}$ values. Cases where $\widehat{R}$ values were smaller than 1.1 were considered good (Gelman & Shirley, 2011). To assess the efficiency of the estimator, we calculated ESS and the Bayesian fraction of missing information (BFMI). Values of ESS above 400 were considered good (Zitzmann & Hecht, 2019), and so were BFMI values above .3 (Betancourt, 2017). Replications where $\widehat{R}$ was larger than 1.1 were excluded from further analyses of bias and coverage.

Furthermore, for each parameter in the model, we calculated the difference between the true and the estimated value and report on the distribution of these values across the 50 replications. Coverage of the parameter estimates was evaluated by examining relative number of replications for which the 95% credibility interval of the posterior distribution includes the true value of the parameter. Following Muthén and Muthén (2002), coverage between.91 and.98 was considered good.

Results

Convergence. Table 2 shows the proportion of analyses in each simulation condition for which all parameter estimates in the model satisfied the criteria of model convergence and estimation efficiency ($\widehat{R} < 1.1$, and $BFMI > .3$). There were no simulation conditions under which all the model parameters had ESS greater than 400.

The results suggest that convergence is good when the variances of the person parameters (specifically of the change in speed and the change in ability) are large. If there are hardly any interindividual differences in intraindividual change, the model is hard to estimate and does not converge. This is not surprising, as the estimation of an effect which does not exist in the data is in general challenging. This is further corroborated by the fact that, on the parameter level, most parameters achieve good convergence and the specific parameter with most convergence issues is the variance of ability change, which also impacts the correlations associated with that parameter (see Figure S5 in the Supplementary materials). As such, convergence issues of change parameters indicate that there are no interindividual differences in change.

While for large variances, convergence of the model is high, a notable exception is the condition with weak correlations, large sample size, and item number. Model estimation is challenging in these conditions as well. Convergence issues specifically

Table 2
Relative Number of Analyses within Each Simulation Conditions that Fulfill the Criteria of $\widehat{R} < 1.1$*, and Bayesian Fraction of Missing Information* $BFMI > .3$

Variance Condition	Correlation Condition	N	K	$\widehat{R} < 1.1$	$BFMI > .3$
Large	Strong	1,500	9	1	1
			25	1	1
		5,000	9	.98	.98
			25	.92	1
		8,500	9	.8	.74
			25	.72	.78
	Weak	1,500	9	.92	1
			25	1	1
		5,000	9	.96	.74
			25	.74	1
		8,500	9	.6	.52
			25	.22	.66
Small	Strong	1,500	9	.16	1
			25	.04	1
		5,000	9	0	.9
			25	.02	1
		8,500	9	.02	.7
			25	0	.88
	Weak	1,500	9	.06	1
			25	0	1
		5,000	9	0	.96
			25	0	1
		8,500	9	0	.62
			25	0	.9

Note: N = sample size, K = number of items, $\widehat{R}$ = R-hat convergence diagnostic, $BFMI$ = Bayesian Fraction of Missing Information.

occur for the correlations of the person parameters and for the variances of the person ability parameters (see Figure S5). It is possible that in larger data sets, due to the larger number of estimated parameters, the currently used number of iterations was not sufficient for proper model estimation.

Because the replications in the condition with small variances displayed low convergence rates, all being even below 20%, we excluded these conditions with small variances from further analyses of bias and coverage.

Assessment of the efficiency of the estimation shows that the sampling process is not optimized. The results show that under no condition an ESS above 400 was obtained for all parameters. However, it is also worth noting that on the parameter level, most of the parameters, including those in the conditions with small variances, had sample sizes over 400 (see Figure S6). Low ESS was obtained for those parameters that also had issues with convergence.

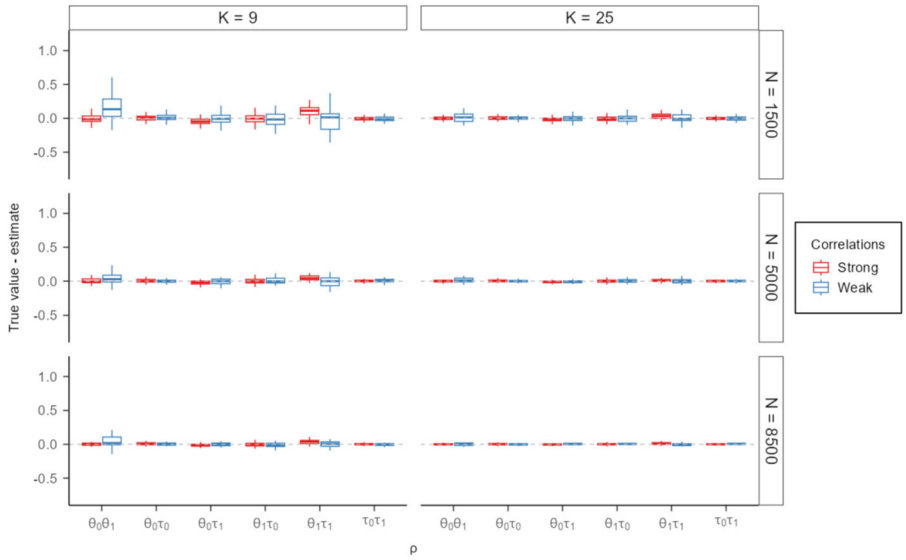

Figure 2. Box-and-whiskers plots showing the difference between the true and the estimated value of the person parameter correlation estimates (bias). Bias was only computed for analyses in which the model converged with $\widehat{R} < 1.1$. K = test length, N = sample size, ρ = correlation, θ_0 = initial ability, θ_1 = ability change, τ_0 = initial speed, τ_1 = speed change. The *x*-axis shows combinations of person parameters for which the correlation was computed.

BFMI shows that the model is efficiently estimated in most conditions, including those with small variances. Several exceptions occur in the conditions with the weak correlations and sample sizes of 8, 500.

Parameter recovery. Figures 2 and 3 display the difference between the true and the estimated value of person parameter correlations and person parameter variances, respectively. Results are only shown for analyses in which the model converged with $\widehat{R} < 1.1$ and are only shown for the simulation study conditions with the large variances. The results suggest that there is overall very little bias in the person parameter correlation (Figure 2) and variance estimates (Figure 3) in conditions with large variances. Slight deviations occur for weak correlations in the condition with low number of item ($K = 9$) and persons ($N = 1, 500$). In this condition, the variance of the change in ability as well as the correlation of initial ability and change in ability are slightly biased. Estimates of the item parameters were unbiased (see Figure S3).

Tables 3 and 4 display the coverage of person parameter correlations and person parameter variances, respectively, for all simulation study conditions with large variances. The coverage of the parameters was very good in most cases, that is, being between.91 and.98. Only in conditions with weak correlation and small number of items, parameters related to the change in ability (θ_1) showed slightly lower coverage rates (the smallest being.83). Coverage of the item parameters was always excellent (between.91 and.98, see Table S1).

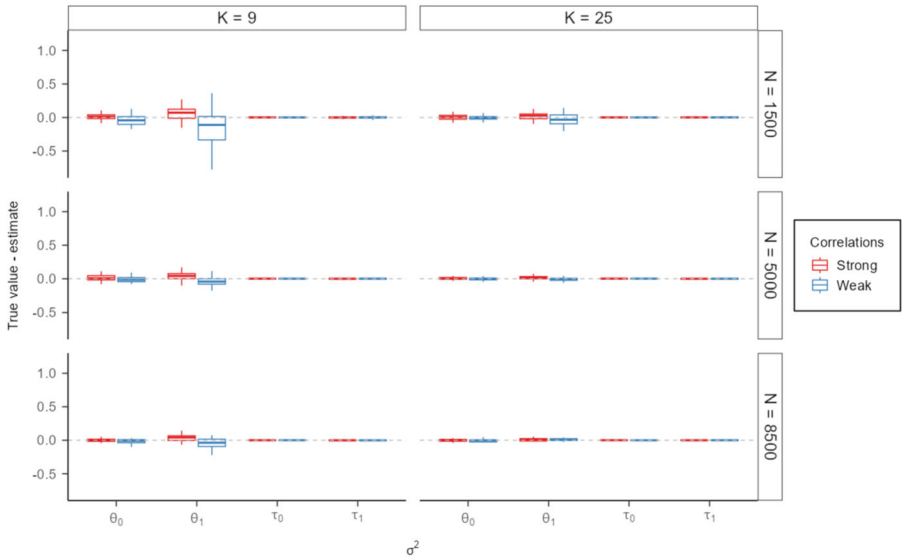

Figure 3. Box-and-whiskers plots showing the difference between the true and the estimated value of the person parameter variance estimates (bias). Bias was only computed for analyses in which the model converged with $\widehat{R} < 1.1$. K = test length, N = sample size, σ^2 = variance, θ_0 = initial ability, θ_1 = ability change, τ_0 = initial speed, τ_1 = speed change.

Table 3
Coverage of the Person Parameter Correlations Using 95% Credibility Intervals of the Posterior Distribution

			ρ					
Correlation Condition	*N*	*K*	$\theta_0\theta_1$	$\theta_0\tau_0$	$\theta_0\tau_1$	$\theta_1\tau_0$	$\theta_1\tau_1$	$\tau_0\tau_1$
Strong	1,500	25	1	1	.94	1	.86	.96
		9	.88	.96	.88	.92	.82	.96
	5,000	25	.98	.93	.96	.93	.96	.96
		9	.94	.92	.92	.9	.82	.96
	8,500	25	.97	.89	.92	1	.89	.97
		9	.98	.93	.9	.9	.85	.9
Weak	1,500	25	.98	.98	.88	.96	.9	.92
		9	.93	.96	.85	.98	.91	1
	5,000	25	.97	.97	.97	.95	.95	.97
		9	.94	.9	.92	.98	.92	.92
	8,500	25	1	1	.91	1	1	1
		9	.83	.97	1	.93	1	.97

Note: N = sample size, K = number of items, ρ = correlation between person parameters. θ and τ represent ability and speed and 0 and 1 in their subscripts represent the initial value and the change, respectively.

Table 4
Coverage of the Person Parameter Variances Using 95% Credibility Intervals of the Posterior Distribution

Correlation Condition	N	K	σ^2			
			θ_0	θ_1	τ_0	τ_1
Strong	1,500	25	.92	.96	.92	.94
		9	.94	.98	.94	.96
	5,000	25	.96	.91	.87	.98
		9	.94	.96	.92	.94
	8,500	25	.94	.94	.97	.97
		9	1	.98	.95	.93
Weak	1,500	25	.94	.92	.9	.92
		9	.93	.87	.93	.93
	5,000	25	.95	1	.92	.92
		9	.96	.96	.92	.96
	8,500	25	.91	1	1	1
		9	.87	.87	.97	.93

Note: N = sample size, K = number of items, σ^2 = variance of person parameters. θ and τ represent ability and speed and 0 and 1 in their subscripts represent the initial value and the change, respectively.

Empirical Example

Data

To illustrate the use of the ISAR model for investigating the intraindividual relation of ability and speed, we made use of data collected as part of the Programme for International Student Assessment (PISA; Organization for Economic Cooperation and Development, 2018) in 2018. We compared two different subject areas, Mathematics and Science, and two different geographical and linguistic regions, the English speaking part of North America[2] and the Spanish speaking part of Latin America.[3] The application of PISA exams in 2018 was hybrid—some examinees solved the exams in a pen-and-paper setting, while the others solved the exams on computers. As we rely on response time data, we only considered individuals in our analyses who solved the exams on computers. This resulted in the following sample sizes included in the analyses: in Latin America 4,612 persons on Mathematics items and 4,390 persons on Science items; in North America 1,998 persons on Mathematics items and 1,875 persons on Science items.

For the analyses, we selected the first cluster of items of the Mathematics and Science test, comprising 12 and 20 items, respectively. Items for which it was possible to achieve partial credit were recoded to binary variables; only if the item was fully solved, it was scored as 1, otherwise it was scores as 0.

Analyses

We applied the ISAR model to each of the four data sets separately. The analyses were run with the same setup as in the simulation study.

To assess the fit of the model, we performed posterior predictive checks separately for responses, for response times, and for the covariances of responses and response times. The details of these analyses can be found in the Section Model fit in the empirical study in the Supplementary materials.

In order to illustrate the added value of the ISAR model, we compare the results to those when analyzing the data with the hierarchical model by van der Linden (2007), that assumes stationarity of ability and speed. In other words, the model does not include growth terms for latent speed and ability, but just single parameters for these constructs. In our implementation, on the lower level, the measurement model for the responses was also modeled with a Rasch model. The measurement model for the response time followed a lognormal distribution, were we also assumed the common variance for all response times. On the higher level, the model includes a multivariate distribution of latent speed and ability. We also compared it to the Rasch model without the response times. We implemented the model in Stan (Stan Development Team, 2021), also retaining all the reparameterizations which we used for our proposed model. The same posterior predictive checks were performed for this model and the proposed model.

Results

Convergence. The results suggest that the model has converged well. The largest $\widehat{R}$ values were 1.03 for the Latin America Mathematics data set, 1.06 for the Latin America Science data set, 1.01 for the North America Mathematics data set, and 1.04 for the North America Science data set.

Estimation efficiency. Evaluation of the estimation efficiency via the ESS shows that the sampling process is not optimized. The lowest values of the ESS were 91.41 for the Latin America Mathematics data set, 114.21 for the Latin America Science data set, 141.47 for the North America Mathematics data set, and 191.88 for the North America Science data set.

Model fit. The results of the posterior predictive checks suggest that the model fits the response time data well, except for the first 5% of the fastest response times (see Figure S1). The model also showed a good fit to the response data (see Figure S2).

The fit of the measurement model for the response times showed very little differences to fit achieved when using van der Linden's (2007) model, thus showing no large advantages of our model in revealing the univariate distributions. The proposed model only yielded very slightly lower residuals of the response times. Likewise, the measurement model for the responses of our model yields very similar fit to the empirical data as the model of van der Linden (2007) and even the Rasch model, without the response times.

Parameter Estimates. Estimated model parameter are shown in Table 5, Figure 4, and Table S2. Both in Mathematics and Science, students in North America scored higher than students in Latin America, as indicated by the lower item difficulty parameters b ($\mu_b = 1.46$ as opposed to $\mu_b = .28$ in Science and $\mu_b = 1.73$ as opposed to $\mu_b = .5$ in Mathematics; see Table S2). On average, the two populations hardly differed in their speed as time intensity parameters β were quite similar in

Table 5
Variances of the Person Parameters

	Mathematics		Science	
Variance	North America	Latin America	North America	Latin America
$\sigma^2_{\theta_0}$	.96	1.25	1.18	1.03
$\sigma^2_{\theta_1}$	1.26	.76	.57	.58
$\sigma^2_{\tau_0}$	.27	.31	.37	.46
$\sigma^2_{\tau_1}$	.34	.31	.49	.66

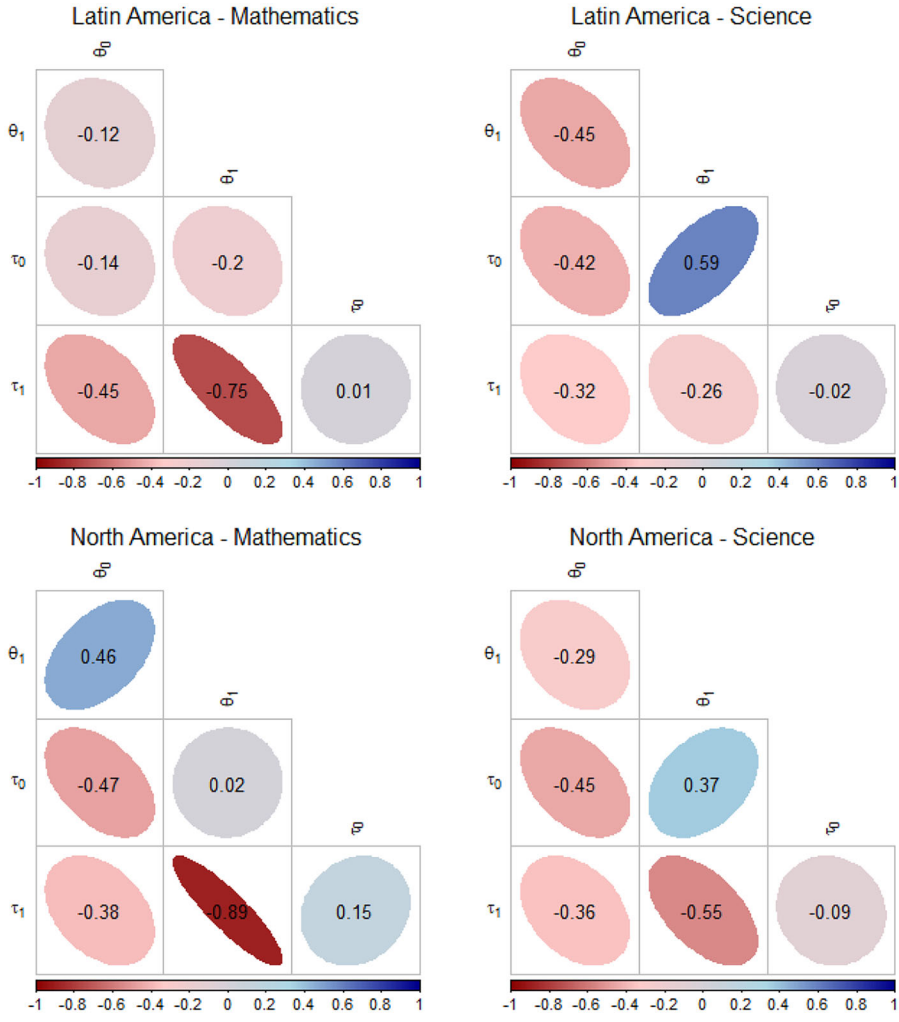

Figure 4. Correlograms of person parameters across the two regions and subject areas.

both groups ($\mu_\beta = 1.9$ and $\mu_\beta = 10.78$ in Science and $\mu_\beta = 11.45$ and $\mu_\beta = 11.23$ in Mathematics).

The variance of the person parameters is shown in Table 5. Students in Latin America were generally more heterogeneous in their initial ability and initial speed than students in North America. There was a high variation in latent change for both speed and ability in all competence domains and regions, supporting the existing of nonstationarity of speed and ability. Thus, person speed and ability level changed across the test. For Mathematics, North Americans showed larger variability in the change of ability across the test than Latin Americans ($\sigma^2_{\theta_1} = 1.26$ as opposed to $\sigma^2_{\theta_1} = .76$), while for the change in speed, parameter estimates were similar ($\sigma^2_{\tau_1} = .34$ and $\sigma^2_{tau_1} = .31$). In Science, the pattern was opposite: for the change in ability, the parameter estimates were similar in North and Latin America ($\sigma^2_{\theta_1} = .57$ and $\sigma^2_{\theta_1} = .58$), while for the change in speed, Latin American examinees showed higher variation ($\sigma^2_{\tau_1} = .66$ as opposed to $\sigma^2_{\tau_1} = .49$).

The estimated correlations of the person parameters are shown in Figure 4. For all data sets, the correlation between initial speed and initial ability was negative, indicating that the faster the students were at the beginning, the lower their effective ability. There was a negative correlation between change in speed and change in ability in all four data sets. Broadly speaking, the negative correlation between these two parameters supports the notion of the existence of the SAT in the data. However, there is considerable variation in this correlation coefficient across the data sets. The highest coefficient ($\rho_{\theta_1\tau_1} = -.89$) was found in the North America - Mathematics data set, while the lowest one ($\rho_{\theta_1\tau_1} = -.26$) was found in the Latin America - Science data set. Both the region and the test subject had an impact on the strength of the correlation: the coefficient was always higher in the Mathematics data sets than in Science data sets, and it was always higher in the North American data sets than Latin American data sets.

Figure 5 shows the intraindividual relation of effective ability and effective speed throughout the test for each person. As we assume a linear change in the model, we only depict two points of this relationship: the initial and the final effective speed and effective ability of each individual. In line with the results on the change in ability and speed, for the majority of individuals, we are able to observe different speed levels, which is one of the prerequisites for investigating intraindividual relations of ability and speed. The results also show that not only the chosen effective speed and the respective effective ability, but also their intraindividual relationship differed considerably across persons. For most of the individuals, the relationship between effective ability and effective speed was negative. This is in line with what we would expect in the presence of an SA trade-off. However, there were also students with a positive intraindividual relationship of effective ability and effective speed. This aligns with what one would expect if there are changes in individuals' concentration during the test. Such positive relationships were found for 11.8% of the students in North America and 27.2% of the students in Latin America in Mathematics, and for 33% of the students in North America, and 47% in Latin America in Science. The percentage of examinees with positive relationships of speed and ability was higher for Science than for Mathematics, and it was also higher for students in Latin America than in North America.

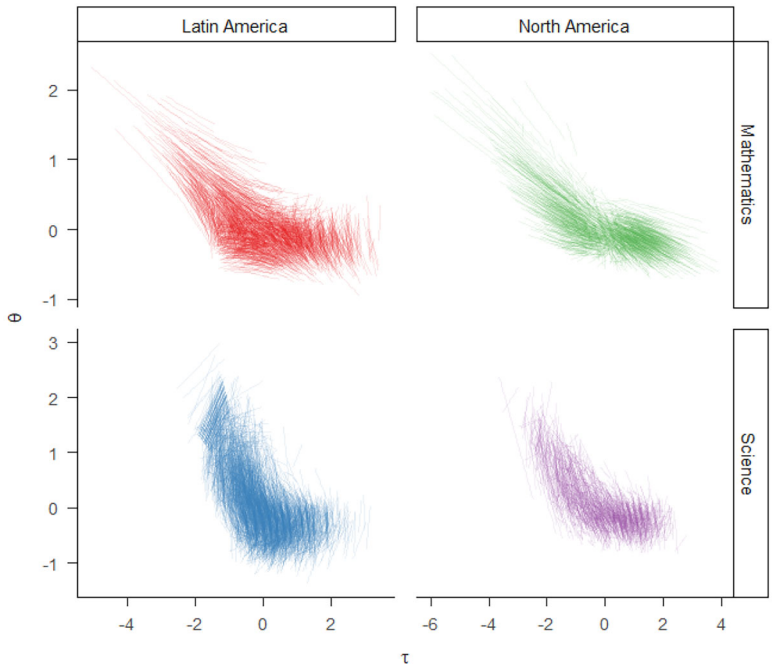

Figure 5. Functions of the speed-ability relationship for every examinee across the two regions and subject areas.

Discussion

In this paper, we provide a model which allows for the explicit examination of the intraindividual relationship between speed and ability in psychometric assessments. By doing so, we add to the canon of approaches that can be used to investigate intraindividual relations of ability and speed. The ISAR model can easily be applied to most psychometric assessments, without the need for experimental manipulation or the availability of external measures. The approach directly models change in speed and ability and, as such, the intraindividual relationship of speed and ability.

We consider our model to be a relevant addition to the literature. As general in assessments, usually only effective ability can be assessed, and examinees do not use the same speed level, which threatens fairness of comparisons across persons. With our approach, we aim to take a step toward depicting the individual SAT of a person, as such getting a more comprehensive picture on the performance of a person. For example, we can investigate the approximated individual SA curve of a person, which may be a better base for comparisons than a single value on effective ability. We advise to take a holistic approach in which both the change in ability and the change in speed are taken into account when considering an individual's performance is preferred to make optimal and fair decisions (Pohl et al., 2021).

The model relies on nonstationarity, that is, that examinees changing their effective speed and their effective ability throughout the test, and, that this change differs between examinees (intraindividual differences in intraindividual change). If there is

no change or examinees do not differ in their change over time, convergence issues occur and the model cannot be estimated. In this case, it is not possible to investigate the intraindividual relation of speed and ability.

For estimating the model, it is required that the variances of the person parameters, specifically of the latent change in ability and speed, are sufficiently large. If there are no interindividual differences in the intraindividual change, parameter estimation becomes challenging. Low convergence in situations with low person parameter variances is, thus, expected, since in these situations there is no need to apply this model. Low convergence of the model may serve as a signal that there is either no change in the ability or speed over the course of the test or that examinees do not differ in their change and that conventional models which assume stationarity, such as the model of van der Linden (2007) can be used instead.

In analyses, in which the model converged, estimates of the distribution of person parameters were largely unbiased. Bias decreases for large sample size and large number of items. This suggests that the model, when it has converged, displays none to minimal bias and can be used to make inferences about the quantities modeled by it.

Convergence and efficiency (i.e., ESS) was lower in conditions with large number of items and specifically persons. In applications with larger number of examinees and items, it may be necessary to increase the number of iterations or chains in the estimation to obtain satisfactory convergence and efficiency.

In the analyses of the empirical data from PISA tests (Organization for Economic Cooperation and Development, 2018), we observed a relationship between effective speed and effective ability that aligns with the interpretation of an SAT for most, but not for all students. While we have found a negative relationship between the change in the latent ability and the change in latent speed in the majority of students, which is the direction that would be expected if the SAT is taking place, there was still a considerable number of cases (11.8–47%, depending on the data set) in which this relation was positive. This could be due to other confounding variables, such as changes in the examinees' concentration during the test (Ranger et al., 2021).

The results of our study align with previous studies, which mostly reported a negative intraindividual relationship between speed and ability, but also identified cases for which the relationship is negative or nonmonotonic (Domingue et al., 2022; Kang et al., 2022; Nietfeld & Bosma, 2003; Ranger et al., 2021).

Extending the model of van der Linden (2007) to include nonstationarity in empirical data was not necessary to improve model fit. As compared to a model assuming stationarity, the ISAR model did only slightly improve model fit. However, parameter estimates showed that there is indeed a considerable amount of nonstationarity in the data (variances of the change rates of ability and speed, $\sigma^2_{\theta_1}$ and $\sigma^2_{\tau_1}$ were nonzero in all of the four analyzed data sets). Thus, change in ability and speed does seem to be present. Most importantly, applying the ISAR model allowed us to investigate the intraindividual relationship of ability and speed for each person, which would not be possible with a model assuming stationarity.

Limitations and Future Directions

Like all models, the ISAR model makes assumptions and has limitations. The model presented in this paper is currently limited to only allow for linear change in effective ability and speed throughout the test. Such a specification may not be an accurate reflection of the process which is taking place in practice. For example, it could be possible that the change in speed happens rather suddenly when the examinee becomes aware that they are running out of time. It may also be that speed is accelerating slowly at the beginning and more steeply at the end of the test. For example, Yamamoto's (1989) HYBRID model assumes a sudden change in speed toward the end of the test. Fox and Marianti's 2016 model assumes a quadratic progression of the latent speed over the course of the test. In order to capture other forms of change, the ISAR model could be extended to incorporate quadratic growth or other nonlinear growth terms. While this is theoretically straightforward, it poses high demands for estimation. Future work may focus on estimation routines for estimating these more complex models.

The ISAR model also makes the assumption that the intraindividual relationship of speed and ability is linear. This is not plausible in all applications. For instance, in the case that there is an SAT, theory (Heitz, 2014; van der Linden, 2009) suggests that the SAT curve is at least quadratic. Goldhammer (2015) describes the shape of the curve as asymptotic, and shows how it can be modeled with a logistic function. It is also possible that the shape of the curve is nonmonotonic (Ranger et al., 2021) as a result of processes of different direction taking place (e.g., the SAT and changes in concentration). While the ISAR model is unable to capture nonlinear relations, (a) it does extend previous approaches (Fox & Marianti, 2016) in that it accounts for the fact that ability may be impacted by speed and (b) allows for examining the direction of the relationship of ability and speed within each individual.

We also made assumptions regarding the measurement models. As previous models (Fox & Marianti, 2016; van der Linden, 2007), we introduced our approach with assuming an IRT model for item responses and a lognormal model of response times. As such, the current formulation of the model could be very sensitive to outliers of response times. Of course, also other measurement models may be incorporated instead. For example, Ranger and Kuhn (2012) implemented a model, which embeds a proportional hazard model and an accelerated failure time model into a link function, for which they binarized response times. It is also possible to specify different measurement models for responses, such as the 2PL (2-parameter logistic) or the 3PL (3-parameter logistic) model. For example, due to the insufficiencies of the Rasch model, OECD switched to the 2PL model (Birnbaum, 1968) and the generalized partial credit model (Muraki, 1992) in 2015 (Organization for Economic Cooperation and Development, 2015).

In settings where items are administered in the same order to all persons, it is not possible to distinguish between item parameters and average change in person parameters. As such, we cannot draw any conclusions on average change in effective ability and speed across the test. However, if items are presented in random order to each person, the item and person effects can be disentangled and average change in

effective ability and speed can be identified. In this situation, instead of fixing the mean of the change variables to zero, the value of the item difficulty and item time intensity of one item may be fixed to a specific value, for instance, to zero, across all positions.

Similar as in other approaches that rely on nonstationarity, the intraindividual relationship may represent the SAT, but also alternative explanations for this relationship exist. As the approach relies on change over the course of the test, other confounding factors, such as learning, motivation, fatigue, or test-taking strategy (Ranger et al., 2021) may change as well. In fact, our empirical results but also those of previous studies (Domingue et al., 2022; Ranger et al., 2021; Kang et al., 2022) suggest that this is indeed the case. To disentangle the impact of the SAT from that of these confounding factors on the intraindividual relationship of ability and speed, one may measure such confounders (e.g., change in motivation or change in concentration) and statistically control for their effects on the change in ability and speed (Much, Mutak, Pohl, & Ranger, 2023).

The proposed approach was presented for investigating the latent change and the intraindividual relationship between ability and speed. The approach may also be applied to other variables than ability and speed, for example, to speed and omission propensity (Glas & Pimentel, 2008; Moustaki & O'Muircheartaigh, 1999; Ulitzsch et al., 2020c) or ability and test-taking engagement (Schnipke & Scrams, 1997; Ulitzsch et al., 2020a). This would not only offer valuable information on the trajectory of omission propensity throughout the course of the test, but also on the intraindividual relation of, for example, speed and omission propensity. This may, in turn, help to understand the process resulting in missing responses.

Acknowledgments

Open access funding enabled and organized by Projekt DEAL.

Notes

[1]We choose this number of iterations per analyses, as in test runs prior to the simulation, we observed a satisfactory model performance with this specification. It is also possible to estimate the model with less but longer chains. The choice of using 10 chains with 10,000 iterations was due to technical requirements of the high-performance cluster that we were using.

[2]Canadian examinees who solved the PISA assessments in French were excluded from the sample.

[3]These included Chile, Columbia, Costa Rica, Dominican Republic, Mexico, Panama, and Peru. Argentina was not included in the sample because response time data for Argentina were not available.

References

Alfers, T., Gittler, G., Ulitzsch, E., & Pohl, S. (2021). Under pressure: Measuring cognitive abilities under instruction-induced time pressure. *6th International NEPS Conference (Virtual Meeting)*.

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. https://doi.org/10.48550/ARXIV.1408.4050

Bennett, L., Melchers, B., & Proppe, B. (2020). *Curta: A General-purpose High-Performance Computer at ZEDAT, Freie Universität Berlin*. https://doi.org/10.17169/refubium-26754

Benoit, E. (1924). Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés a un système d'équations linéaires en nombre inférieur a celui des inconnues. - application de la méthode a la résolution d'un système defini d'équations liné aires. *Bulletin Géodésique*, *2*(1), 67–77. https://doi.org/10.1007/BF03031308

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. https://doi.org/10.48550/ARXIV.1701.02434

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.), *Current trends in Bayesian methodology with applications* (pp. 79–102). Boca Raton, FL: CRC Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord& M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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., & 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. https://doi.org/10.3389/fpsyg.2017.00202

Domingue, B. W., Kanopka, K., Stenhaug, B., Sulik, M. J., Beverly, T., Brinkhuis, M., Circi, R., Faul, J., Liao, D., McCandliss, B., Obradović, J., Piech, C., Porter, T., Project iLEAD Consortium, Soland, J., Weeks, J., Wise, S., & Yeatman, J. (2022). Speed accuracy tradeoff? Not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real-world settings. *Journal of Educational and Behavioral Statistics*, *47*(5), 576–602. https://doi.org/10.3102/10769986221099906

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 440–553. https://doi.org/10.1080/00273171.2016.1171128

Fox, J. (2020). *Regression diagnostics: An introduction*. Thousand Oaks, CA: Sage Publishing.

Frey, A., Spoden, C., Goldhammer, F., & Wenzel, S. F. C. (2018). Response time-based treatment of omitted responses in computer-based testing. *Behaviormetrika*, *45*, 505–526. https://doi.org/10.1007/s41237-018-0073-9

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, *71*(2), 369–382. https://doi.org/10.1111/j.1751-5823.2003.tb00203.x

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, *13*(4), 755–779. https://doi.org/10.1198/106186004X11435

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis*. New York: Chapman and Hall/CRC.

Gelman, A., Goegebeur, Y., Tuerlickx, F., & Mechelen, I. V. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of Applied Statistics*, *49*(2), 247–268. https://doi.org/10.1111/1467-9876.00190

Gelman, A., Jakulin, A., Pittau, M., & Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*, 1360–1383.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 163–174). Boca Raton, FL: CRC Press.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907–922. https://doi.org/10.1177/0013164408315262

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*, 133–164. https://doi.org/10.1080/15366367.2015.1100020

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766

Guo, X., Luo, Z., & Yu, X. (2020). A speed-accuracy tradeoff hierarchical model based on cognitive experiment. *Frontiers in Psychology*, *10*, 2910. https://doi.org/10.3389/fpsyg.2019.02910

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150. https://doi.org/10.3389/fnins.2014.00150

Higham, N. J. (2009). Cholesky factorization. *Wiley Interdisciplinary Reviews Computational Statistics*, *1*, 251–254. https://doi.org/10.1002/wics.18

Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression*. New York: Wiley.

Kang, I., De Boeck, P., & Ratcliff, R. (2022). Modeling conditional dependence of response accuracy and response time with the Diffusion Item Response Theory model. *Psychometrika*, *87*(2), 725–748. https://doi.org/10.1007/s11336-021-09819-5

Krause, R., Mutak, A., Much, S., Alfers, T., Ulitzsch, E., Ranger, J., & Pohl, S. (2022). *Aiming at identifying the speed-accuracy-tradeoff in psychological tests by modeling conditional dependence of responses and response times*. (No. 52). Hildesheim, Germany: Kongress der Deutschen Gesellschaft für Psychologie.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 14. https://doi.org/10.3390/jintelligence4040014

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Liu, 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

Liu, J., Wang, C., & Shi, N. (2023). A mixture response time process model for aberrant behaviors and item nonresponses. *Multivariate Behavioral Research*, *58*(1), 71–89. https://doi.org/10.1080/00273171.2021.1948815

Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, *43*(8), 639–654. https://doi.org/10.1177/0146621618824853

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. https://doi.org/10.1111/jedm.12060

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205–228. https://doi.org/10.1111/bmsp.12117

Molenaar, D., Oberski, D., Vermunt, J. K., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626. https://doi.org/10.1080/00273171.2016.1192983

Molenaar, D., Tuerlinckx, F., & van der Maas, H. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219. https://doi.org/10.1111/bmsp.12042

Moustaki, I., & O'Muircheartaigh, C. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society*, *162*(2), 177–194. https://doi.org/10.1111/1467-985X.00129

Much, S., Mutak, A., Pohl, S., & Ranger, J. (2023). Modeling speed-ability trade-off and test-taking persistence - parameter validation for two psychometric models. https://doi.org/10.17605/OSF.IO/9J6HM

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–177. https://doi.org/10.1177/014662169201600206

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*(3), 705–767. https://doi.org/10.1214/aos/1056562461

Nietfeld, J., & Bosma, A. (2003). Examining the self-regulation of impulsive and reflective response styles on academic tasks. *Journal of Research in Personality*, *37*(3), 118–140. https://doi.org/10.1016/S0092-6566(02)00564-0

Organization for Economic Cooperation and Development. (2015). *PISA 2015 technical report*. Technical Report.

Organization for Economic Cooperation and Development. (2018). *PISA 2018 technical report*. Technical Report.

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, *22*(1), 59–73. https://doi.org/10.1214/088342307000000014

Peng, S., Cai, Y., Wang, D., Luo, F., & Tu, D. (2022). A generalized diagnostic classification modeling framework integrating differential speediness: Advantages and illustrations in psychological and educational testing. *Multivariate Behavioral Research*, *57*(6), 940–959. https://doi.org/10.1080/00273171.2021.1928474

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. https://doi.org/10.1126/science.abd3300

Ranger, J., & Kuhn, J.-T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31–47.

Ranger, J., Kuhn, J. T., & Pohl, S. (2021). Effects of motivation on the accuracy and speed of responding in tests: The speed-accuracy tradeoff revisited. *Measurement: Interdisciplinary Research and Perspectives*, *19*(1), 15–38. https://doi.org/10.1080/15366367.2020.1750934

Scharfen, J., Blum, D., & Holling, H. (2018). Response time reduction due to retesting in mental speed tests: A meta-analysis. *Journal of Intelligence*, *6*(1), 6. https://doi.org/10.3390/jintelligence6010006

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Stan Development Team. (2021). *Stan modeling language users guide and reference manual*. https://mc-stan.org

Ulitzsch, E., von Davier, M., & Pohl, S. (2020a). 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*, *83*(S1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., von Davier, M., & Pohl, S. (2020b). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*, *80*(3), 522–547. https://doi.org/10.1177/0013164419878241

Ulitzsch, E., von Davier, M., & Pohl, S. (2020c). 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 Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, *70*(2), 359–376. https://doi.org/10.1007/s11336-003-1078-0

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20. https://doi.org/10.3102/1076998607302626

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/S11336-007-9046-8

van der Maas, H. L. J., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*(1), 29–60. https://doi.org/10.2307/30039042

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. https://doi.org/10.1111/bmsp.12054

Wise, S. I., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (ETS Research Report RR-89-41).

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 646–661. https://doi.org/10.1080/10705511.2018.1545232

Appendix: Stan Code

```
data {
  int<lower=0> n_person; // number of examinees
  int<lower=0> n_item; // number of items
  int<lower = 1> Dim; // number of dimensions
  int<lower = 1> N; // number of data points
  int<lower = 1> jj[N]; // item id
  int<lower = 1> ii[N]; // person id
  int<lower=0, upper = 1> Y[N];
  vector<lower=0>[N] t; // response time data
  vector<lower=0, upper=1>[N] X; // time scale variable
  vector[Dim] Zero; // vector of 0s for person parameter means
  }

parameters {
  //relevant parameters
  vector[n_item] b; // item difficulty parameters
  vector[n_item] beta; // item time intensity parameters
  cholesky_factor_corr[Dim] choleskyP; // Cholesky decomposition of correlation of person
      parameters
  //provisional variables for reparametrizations
  matrix[n_person, Dim] PersParStar;
  vector<lower=0,upper=pi()/2>[Dim] sigmaPStar;
  real<lower=0,upper=pi()/2> sigmaTStar;
}

transformed parameters {
  vector<lower=0>[Dim] sigmaP = 2.5 * tan(sigmaPStar); //person parameter variances
  real<lower=0> sigmaT = 2.5 * tan(sigmaTStar); //common variance of RTs
  matrix[n_person, Dim] PersPar = PersParStar * diag_pre_multiply(sigmaP, choleskyP)'; //
      person parameters
}

model {
  to_vector(PersParStar) ~ std_normal();
  choleskyP ~ lkj_corr_cholesky(1);
  b ~ normal(0, 2);
  beta ~ normal(0, 2);
  sigmaPStar ~ uniform(0, pi()/2);
  sigmaTStar ~ uniform(0, pi()/2);
  target += bernoulli_logit_lpmf(Y | (PersPar[ii,1] + PersPar[ii,2] .* X) - b[jj]);
  target += lognormal_lpdf(t | beta[jj] - (PersPar[ii,3] + PersPar[ii,4] .* X), sigmaT);
}

generated quantities {
  matrix[Dim, Dim] correlP = multiply_lower_tri_self_transpose(choleskyP); //correlation
      matrix of person parameters
  matrix[Dim, Dim] SigmaP = quad_form_diag(correlP, sigmaP); //variance-covariance matrix
      of person parameters
  int<lower=0, upper = 1> Y_rep[N] = bernoulli_logit_rng((PersPar[ii,1] + PersPar[ii,2]
      .* X) - b[jj]);
  real<lower=0> t_rep[N] = lognormal_rng(beta[jj] - (PersPar[ii,3] + PersPar[ii,4] .* X),
      sigmaT);
}
```

Authors

AUGUSTIN MUTAK is a doctoral student and a research assistant at Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany; augustin.mutak@fu-berlin.de. His primary research interests include item response theory and affect rhythmicity.

ROBERT KRAUSE is an Assistant Professor at University of Kentucky, 550 South Limestone, Lexington, KY 40508; robert.w.krause@uky.edu. His primary research interests include social network analysis, social influence, and cognitive social structures.

ESTHER ULITZSCH is an Associate Professor at the Centre for Educational Measurement (CEMO) at the University of Oslo, Gaustadalleen 21, 0373 Oslo, Norway;

ulitzsch@cemo.uio.no. Her primary research interests include psychometric modeling, process data analysis, and estimation of latent variable models.

SÖREN MUCH is a research assistant at Martin-Luther-Universität Halle-Wittenberg, Emil-Abderhalden-Str. 26-27, 06108 Halle, Germany; soeren.much@psych.uni-halle.de and a doctoral student at Freie Universität Berlin. His primary research interests include psychometric evidence accumulation models and test-taking behavior.

JOCHEN RANGER is a research assistant at Martin-Luther-Universität Halle-Wittenberg, Emil-Abderhalden-Str. 26-27, 06108 Halle, Germany; jochen.ranger@psych.uni-halle.de. His research is focused on response time modeling in tests and the analysis of model fit.

STEFFI POHL is a professor for methods and evaluation/quality assurance at Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany; steffi.pohl@fu-berlin.de. Her primary research interests include psychometric modeling and causal inference.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Response time fit diagnostics for the ISAR model and van der Linden's (2007) model (SA).
Figure S2. Response fit diagnostics for the ISAR model, van der Linden's (2007) model (SA) and the Rasch model.
Figure S4. Person parameter distributions across the two regions and subject areas.
Figure S3. Box-and-whiskers plots showing the difference between the true value and the estimate of the item parameter estimates.
Figure S5. Parameter-specific convergence rates across different conditions of the simulation study.
Figure S6. Parameter-specific effective sample size across different conditions of the simulation study.
Table S1. Coverage of the item parameters using 95% credibility intervals of the posterior distribution.
Table S2. Item difficulty (b) and time intensity (β) parameter estimates in the empirical study.
Data S1