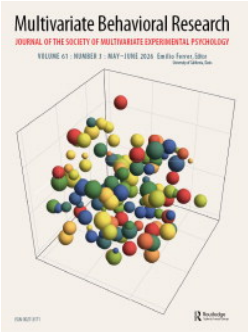

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: www.tandfonline.com/journals/hmbr20

A Unified Framework for Jointly modelling Response Times and Item Position Effects in Computer-Based Learning Assessments

Silvia Bacci, Rosa Fabbricatore & Maria Iannario

To cite this article: Silvia Bacci, Rosa Fabbricatore & Maria Iannario (27 Jun 2026): A Unified Framework for Jointly modelling Response Times and Item Position Effects in Computer-Based Learning Assessments, Multivariate Behavioral Research, DOI: 10.1080/00273171.2026.2684069

To link to this article: https://doi.org/10.1080/00273171.2026.2684069

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2026.2684069

Routledge
Taylor & Francis Group

OPEN ACCESS
Check for updates

A Unified Framework for Jointly modelling Response Times and Item Position Effects in Computer-Based Learning Assessments

Silvia Bacci[a], Rosa Fabbricatore[b], and Maria Iannario[c]

[a]Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Firenze, Italy; [b]Department of Economics and Statistics, University of Naples Federico II, Napoli, Italy; [c]CA: Department of Political Sciences, University of Naples Federico II, Napoli, Italy

ABSTRACT

Current models for assessing response accuracy and response times in *testing environments* typically overlook variations in speed and ability within individuals. Instead, they often treat these as residual variances, missing the dynamic changes in individual performance throughout a test. Additionally, the influence of item position and the ordinal nature of response accuracy remain underexplored. This paper introduces a comprehensive modelling framework that integrates item responses, response times, and item position to better understand skill acquisition and latent speed changes. Our approach uses the Bivariate Generalised Linear Item Response Theory (B-GLIRT) model, capturing the dual impact of ability on response accuracy and the interplay between ability and speed on response times. We extend this model by incorporating random effects for item and individual-specific variations. The proposed model further explores how item positioning affects test performance and provides diagnostic insights into individual differences. The paper also discusses parameter estimation, model identification, and applications to real-world data, illustrating the practical implications of our findings in computer-based learning assessments.

KEYWORDS

B-GLIRT model; educational assessment; item position; item response theory; latent traits; measurement invariance; ordinal data; response times

Introduction

Inference on individual differences in test responses is traditionally based on latent variable analysis. It deals with the specification of appropriate measurement models to link the observed item scores to the underlying latent variable the test intends to measure. Most measurement models are special cases of a general class of models commonly referred to as *Generalised Linear Item Response Theory* (GLIRT; Bartholomew et al., 2011; Mellenbergh, 1994; Moustaki & Knott, 2000; Skrondal & Rabe-Hesketh, 2004). All models within GLIRT have been developed with the specific objective of analysing the responses of the items collected using traditional *paper* and *pencil* tests.

Item responses are increasingly collected through computerised tests, providing not only item scores but also response times. These response times offer valuable insights into latent ability, particularly through the 'speed-accuracy trade-off': faster responses may reflect less thoughtful consideration. However, recent research challenges the conventional view. In certain domains, such as computer-based mathematics assessments, a negative speed-accuracy trade-off is observed. In these contexts, higher-ability respondents tend to spend more time on test items, while lower-ability respondents are quicker but less accurate (Bungaro et al., 2024). This finding reinforces the notion that in task domains requiring deliberate, controlled cognitive processing, increased time on task is often linked to improved performance, as test-takers engage in more thorough reasoning and problem-solving (Goldhammer et al., 2014; Naumann & Goldhammer, 2017).

Jointly modelling item responses and response times can enhance the precision of latent trait estimation (e.g., Ranger & Ortner, 2011; Van der Linden et al., 2010), although the gains are often modest unless specific conditions are met, such as with short tests or at extreme levels of ability (Ranger, 2013). Additionally, this approach supports test development by improving aspects such as item calibration and item selection in adaptive testing (van der Linden, 2007).

CONTACT Maria Iannario maria.iannario@unina.it Department of Political Sciences, University of Naples Federico II, 80133 Napoli, Italy
Supplemental data for this article is available online at https://doi.org/10.1080/00273171.2026.2684069.

One further issue in achievement testing is administering the same set of items in different orders to prevent copying and enhance test security. Modelling item-position effects within an Item Response Theory (IRT) framework has been extensively addressed. Among others, see Debeer and Janssen (2013) and Wu et al. (2019), who consider descriptive and explanatory IRT models; De Boeck and Wilson (2004) to model item position effect in terms of Differential Item Functioning (DIF; Holland & Wainer, 1993); Albano (2013) who relies on a multilevel modelling of item position effects, and Weirich et al. (2014) who refer to the class of Generalised Linear Mixed Models (GLMMs; Clayton, 1996; McCulloch & Searle, 2001; Skrondal & Rabe-Hesketh, 2004), also known as hierarchical generalised linear models (e.g., Raudenbush & Bryk, 2002), for the position effect. The latter is also the reference approach for the present proposal. Specifically, we extend this line of research by embedding item position modelling within a structural equation modeling (SEM; Bollen et al., 2008) framework that simultaneously captures the dynamics of response accuracy and response times, as detailed below.

Problem statement and contribution

In the context mentioned above, apart from the work of Schweizer and Ren (2013), who discuss the effect of item position constrained in terms of timing, the relationship between time, item position, response accuracy, and respondents' abilities has been marginally explored, mainly when assessed in terms of respondent profiling.

The present paper aims to contribute to this research by proposing a solution that considers *response time* data available at the item level and *item location* effect when the same items are presented in different positions within multiple test forms. Furthermore, the approach refers to answers structured on an *ordinal scale*, where the order represents the level of accuracy of the answer, leading to a different approach concerning the classical setup of binary responses (correct/incorrect). The model can be framed within the class of *Bivariate Generalised Linear Item Response Theory* models (B-GLIRT; Molenaar et al., 2015) that base the assessment of students' ability on jointly considered item responses and response times. The framework is extended in the GLIRT component by incorporating item positioning and random effects into the predictor (Rijmen et al., 2003). This enhancement integrates interaction- and item-dependent effects, drawing inspiration from the structure of IRTree models (De Boeck & Partchev, 2012). The core of our contribution lies in the statistical modelling and inference innovations we introduce. These developments are directly relevant to psychometric and behavioural theory. In the main text, we illustrate the proposed framework through a case study situated in the educational domain. Moreover, to demonstrate the broader applicability of the B-GLIRT approach beyond educational assessment, the Supplementary Material presents an application to a behavioural research problem, namely individuals' propensity to attribute sex to visual stimuli.

The main research questions concern:

1. How the position of an item affects:
   i. response accuracy (i.e., the degree to which a given response correctly aligns with a predefined standard, reference, or "true" value),
   ii. response time;
2. To what extent does item position interact with individual latent traits, such as ability and speediness;
3. How individual characteristics (e.g., gender) affect the model's measurement and structural components.

In sum, jointly modelling item responses expressed on an ordinal scale of correctness and response times within a GLIRT framework, while also accounting for item position effects, constitutes a novel methodological advancement in latent variable modelling. This integrated approach allows for a more nuanced understanding of individual differences, accounting for multidimensional sources of variation in test performance. As such, it represents a significant contribution to ongoing developments in psychometrics and multivariate modelling.

The remainder of the paper is organised as follows. In the section "Methodological background" we present a review of the main literature on modelling speed-accuracy and item position effect. In section "The B-GLIRT model", the general formulation of B-GLIRT models is provided, with details about different forms of the cross-relation between speed and ability, model inference, measurement invariance, and identifiability. Furthermore, the section "Case study" discusses an application to real-world data concerning computerised tests submitted in an Italian higher education institution. A concluding remarks section ends the paper.

Methodological background

Issues related to integrating item response data with response times on one hand and with item position effects on the other have been extensively discussed in the psychometric and statistical literature. In what follows, we first outline the main approaches for modelling item responses and response times; then, we describe the literature proposals for assessing item position effects.

Speed-ability measurement models

In light of the increasing diffusion of computer-based tests, various proposals have been made in the recent literature to incorporate response times as an additional source of information in the analysis of item responses. A comprehensive overview of the main approaches is provided in De Boeck and Jeon (2019). They distinguish three types of models: (i) response time models, where the focus is on modelling the response times, which are the unique response variable of interest; (ii) models where the focus is on explaining another variable (e.g., latent ability), with the response times playing the role of an independent variable; (iii) joint models where the interest is focused to the same extent on two latent traits: the ability measured by the test items and the speed measured by the times recorded for each item response.

Among the joint models, the B-GLIRT models (Molenaar et al., 2015) represent a broad and flexible model class that simultaneously accounts for both response accuracy and response time in testing contexts. These models are particularly suited for understanding the interplay between latent traits, such as ability and speed, by jointly modelling the probability of a correct response and the time taken to provide that response. The B-GLIRT framework allows for a nuanced analysis of how various item and individual-level factors influence test performance, providing insights into complex interactions and dependencies between ability and speed.

The two measurement models are connected with a cross-relation function that may have several specifications. Popular examples of B-GLIRT models to measure ability and speed are the ability model of Thissen (1983) and the hierarchical model of van der Linden (2007).

The primary advantage of using a B-GLIRT model is the possibility of strengthening the measurement ability and achieving higher precision in estimates. Moreover, the high flexibility of generalised linear latent variable models makes it possible to extend the model to account for additional elements, such as the presence of multiple latent traits (e.g., various abilities measured by the test) as well as the influence of individual and item characteristics, including the effect of item position, as specified in more detail below.

Item-position effect models

Issues regarding the ordering of items in a test and the effects on item difficulty, the estimation of latent ability, and, ultimately, the fairness of the test due to multiple test forms with different item orderings have been extensively discussed in the educational and psychological literature. For an overview, refer to Leary and Dorans (1985) for early contributions and Wang (2019) for more recent ones. Generally, the literature outlines two main conflicting effects, depending on the methodological approach and contextual setting: a fatigue effect, where item difficulty increases with item position (i.e., as the item approaches the end of the test), and a practice or learning effect, which operates in the opposite direction (i.e., item difficulty decreases as item position increases).

The methodological approaches adopted in the literature to model item position effects can be categorised into different streams. One approach involves formulating a separate logistic model for each (binary) item, with the linear predictor depending on the raw score of the test and the item position. The item position may be included in the model as either a categorical variable (Pomplun & Ritchie, 2004) or a continuous variable (Davey & Lee, 2011). However, this approach overlooks the latent nature of the ability measured by the test.

This limitation is addressed by approaches based on IRT, where the contribution of each item to assess latent ability is considered within a global framework. In such a setting, Hartig and Buchholz (2012), Debeer and Janssen (2013), Albano (2013), and Weirich et al. (2014) were among the first to advance proposals to treat the item position; see also Alexandrowicz and Matschinger (2008) for a simulation study concerning the robustness of the GLMM parametrisation when the assumption of normality of the latent ability is violated.

Models formulated in the framework at issue resulted in explanatory IRT (De Boeck & Wilson, 2004), or latent regression IRT models (Mislevy, 1985, 1987; Moustaki & Knott, 2000; Zwinderman, 1991), with item position entering the linear predictor as an independent continuous or categorical variable. Namely, a linear effect is assumed when a continuous variable is used, whereas a non-linear effect is specified using a categorical variable. Besides, this effect may be item-constant (i.e., the impact of position does not depend on the item) or item-specific (i.e., the effect of position varies with the item).

Hartig and Buchholz (2012) and Debeer and Janssen (2013) also introduced a random effect to account for a person-specific effect of item position, thus identifying a further latent variable, commonly named *persistence*. The resulting models are configured as a specification of a two-dimensional IRT model (Reckase, 2009), with the two latent variables (ability and persistence) possibly correlated. Trendtel and Robitzsch (2018) and Trendtel and Robitzsch (2021) extended the model of Debeer and Janssen (2013) to the Bayesian setting, allowing for the presence of item position effects on both the item and person sides within the same model. Bulut et al. (2017) presented an alternative model specification that allows for a person-specific effect (Bulut et al., 2017). Within the structural equation modelling setting (Bollen et al., 2008; Duncan, 1975; Goldberger, 1972), they formulated a latent regression unidimensional IRT model with an interaction term between latent ability and item position in the linear predictor (in addition to a linear item-specific position effect).

In the mentioned contributions, the effect of item position typically enters the model additively, thus adding to the item difficulty. However, as outlined by Nagy et al. (2018), the effect of item position may also impact the item discrimination power, in which case it enters the model multiplicatively (i.e., as a multiplicative coefficient of the item discrimination parameters). These two types of effect, one on item difficulty and the other on item discrimination may be interpreted in terms of uniform DIF and non-uniform DIF, respectively (Debeer & Janssen, 2013).

The GLMM framework is particularly suitable for expanding the IRT model by incorporating individual and item characteristics that may serve as mediators regarding the effect of item position and having a direct effect on ability and persistence. For examples see, among others, Nagy et al. (2018), Wu et al. (2019) and Ong and Pastor (2022).

Joint analysis of response times and item position effect

To conclude this review, we note that the two streams of literature discussed above have largely evolved in parallel, with only limited points of intersection. Three notable exceptions are the contributions of Li et al. (2012), Vida et al. (2021), and Becker et al. (2022).

Li et al. (2012) and Vida et al. (2021) formulated separate models to analyse the effect of item position on both item difficulty and response time. Rasch-type models were adopted to explain item responses and log-normal models to explain the response times, with item position included as a fixed or a random effect. To our knowledge, these are the first works that simultaneously analyse issues related to response times and item position, albeit using independent models. Li et al. (2012) found only weak and negligible evidence of a position effect in all their models about the ability, speed estimates, and item parameters. Also, Vida et al. (2021) confirmed the absence of a significant relation between response accuracy and item position. At the same time, they found a statistically significant positive relation between response time and item position (i.e., response times decrease as the item is positioned later in the test).

In contrast, Becker et al. (2022) developed a joint model combining a Rasch model for ability and a log-normal model for speed, conducting simulation studies to explore the impact of the item position on ability estimates. They found a significant effect of item ordering on test scores, mainly when items were ordered from most difficult to easiest. It is worth noting that the joint model of Becker et al. (2022) does not directly model the item position effect.

In the present contribution, we propose formulating a joint model belonging to the B-GLIRT class to measure both ability and speed simultaneously while testing the effect of item position on response accuracy and response time.

The B-GLIRT model

In what follows, we first illustrate the specification of the class of B-GLIRT models. Then, we focus on the introduction of interaction effects and the inspection of the measurement invariance. We also provide some hints about model estimation and model identifiability.

Model formulation

For a sample of N individuals and J items, we assume that two underlying latent variables drive the response process of individual i measured by the test items:

- the first latent variable Θ_i represents the *ability* of each respondent;
- the second latent variable η_i refers to the *speediness* of each respondent.

Let Y_{ij} be the response of individual i to the ordinal item j (*item response*) and T_{ij} the time taken by individual i to answer item j (*response time*). Both Y_{ij} and T_{ij} derive from two continuous unobservable variables, Y^*_{ij} and T^*_{ij}, where:

$$Y_{ij} = k \qquad \Longleftrightarrow \qquad \tau_{k-1} < Y^*_{ij} \le \tau_k, \quad k = 1, 2, ..., K,$$

and

$$T_{ij} = \exp\left(T^*_{ij}\right).$$

Here $-\infty = \tau_0 < \tau_1 < ... < \tau_K = +\infty$ are the thresholds of the support of Y^*_{ij}. The choice of K determines in how many intervals the support of Y^*_{ij} is divided, and it is strictly related to the number of categories assumed for the responses to each item. In the following, we adopt a probit link to connect the observed ordinal variable Y_{ij} to its latent continuous counterpart Y^*_{ij} and a log link to connect T_{ij} to its latent counterpart T^*_{ij}, under the assumption of standard normal distribution for the latent variables Θ_i and η_i.

We assume that the *item responses* are directly affected by Θ_i, whereas the *response times* depend on the ability Θ_i and the speediness η_i. The path diagram in Figure 1 illustrates the relations among the variables. The resulting model configures as a *within-item bidimensional* model (Adams et al., 1997). It is worth noting that no double arrow links the two latent variables Θ_i and η_i, as the reciprocal relationship between them is captured through a cross-effect represented by the presence of arrows from Θ_i to time indicators T^*_{ij}.

Furthermore, the latent variables Y^*_{ij} and T^*_{ij} may depend on one or more observed covariates. In particular, we assume that the *item positioning* P_{ij} (i.e., the position of item j in the test administered to individual i) could affect both item responses Y^*_{ij} and response times T^*_{ij}. Thus, the path diagram of Figure 1 is modified as shown in Figure 2.

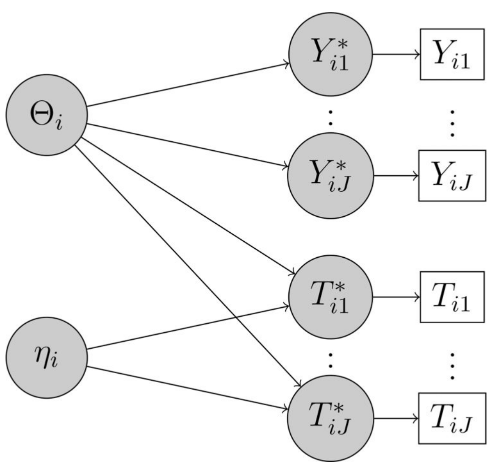

Figure 1. Path diagram of the specified B-GLIRT model. The dependency of item responses and response times on the two latent variables is represented with arrows.

Summarising, the model we propose includes: i) a measurement model linking the item responses to the latent ability variable Θ_i; ii) a measurement model linking the response times to the latent speed variable η_i; iii) a cross-relation function $f(\Theta_i; \rho)$ that connects the two measurement models, being ρ the cross-relation (regression-type) parameter linking ability to response times; (iv) an item-specific effect due to item position P_{ij} that possibly interacts with the latent traits Θ_i and η_i.

Accordingly, the general model we propose is given by:

$$\begin{aligned} \mathbb{E}\left[Y^*_{ij}\right] &= \tau_{jk} + \alpha_j\Theta_i + \omega_{1j}P_{ij} + \omega_{2j}(\Theta_i P_{ij}) \\ \mathbb{E}\left[T^*_{ij}\right] &= \lambda_j + \gamma_j\eta_i + f(\Theta_i; \rho) + \zeta_{1j}P_{ij} + \zeta_{2j}(\eta_i P_{ij}) \end{aligned} \tag{1}$$

with:

- τ_{jk}: item-step parameter denoting the difficulty of selecting category k of item j ($k = 2, ..., K, j = 1, ..., J$);
- λ_j: intercept of the measurement model of the response time;
- α_j and γ_j: factor loading parameters describing the discrimination capability of item j in the assessment of Θ_i and η_i, respectively. Note that when the sign of γ_j is positive, the meaning of the η_i is reversed—lower scores indicate higher speediness—and the

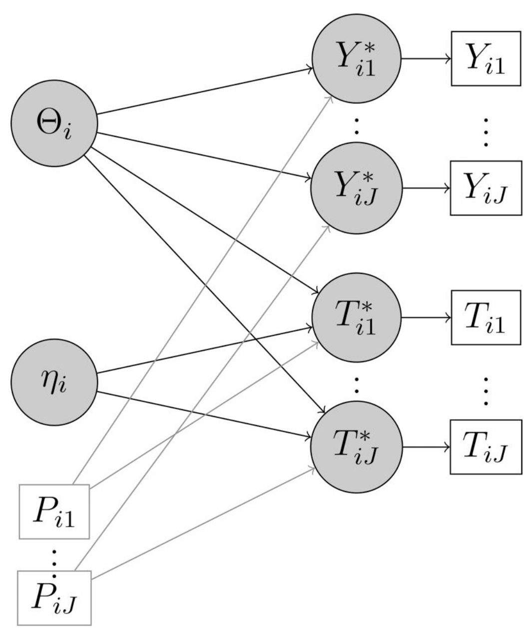

Figure 2. Path diagram of the specified B-GLIRT model, including the position effect. The dependency of item responses and response times on the two latent variables and the item positioning effect is represented with arrows.

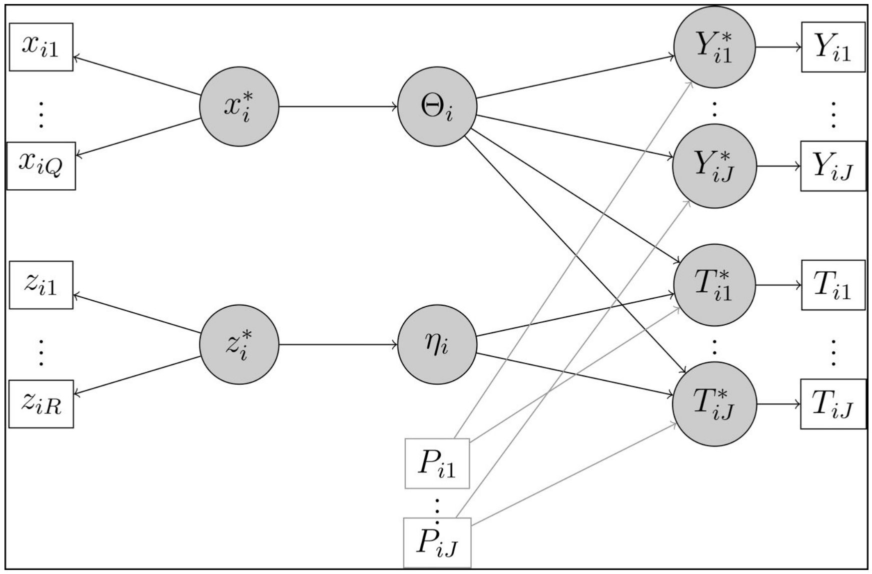

Figure 3. Path diagram of the latent regression B-GLIRT model, with one latent covariate x_i^* on Θ_i and one latent covariate z_i^* on η_i.

association between Θ_i and η_i is expected to be negative;

- ω_{1j} and ζ_{1j}: item-specific linear effect of item position affecting the item response and the item response time, respectively. In other words, we are assuming that the effect of position is specific for each item;
- ω_{2j} and ζ_{2j}: item-specific interaction effects between item position and latent trait (i.e., ability and speediness, respectively). Thus, we assume that the effect of position changes with the latent trait level (and it is item-specific).

Setting suitable constraints, we have:

- $\omega_{1j} = \omega_1$ and/or $\zeta_{1j} = \zeta_1$, for all j: item-constant linear effect of item position. Under this constraint, the effect of position is the same across all items. Similarly, we may set $\omega_{2j} = \omega_2$ and/or $\zeta_{2j} = \zeta_2$, for all j.
- $\omega_{2j} = 0$ and/or $\zeta_{2j} = 0$, for all j: absence of interaction effects between item position and latent trait. Under this constraint, the impact of position does not depend on the latent trait level.

Considering this more general formulation, Figure 2 is modified with two additional latent traits (circles) for each item, denoting the interaction effects that point to Y_{ij}^* and T_{ij}^*, respectively; see also Figure 1 in Bulut et al. (2017) for an illustrative example.

Regarding the cross-relation function, when a *linear form* of $f(\Theta_i; \rho)$ is assumed, two popular models can be reformulated within B-GLIRT (Molenaar et al., 2015): the hierarchical model of van der Linden (2007)

$$f(\Theta_i; \rho) = -\gamma_j \rho \Theta_i \tag{2}$$

and the ability model of Thissen (1983)

$$f(\Theta_i; \rho) = -\alpha_j \rho \Theta_i. \tag{3}$$

It is worth noting that these two formulations represent alternative cross-relation structures; that is, they should not be interpreted as nested components of the same model.

The structural part of the model proposed here may be enhanced to allow for individual characteristics that may directly affect the latent traits by substituting Θ_i and η_i in eq. (1) with

$$\begin{aligned} \Theta_i &= \boldsymbol{x}_i^{*'} \boldsymbol{v} + \varepsilon_i \\ \eta_i &= \boldsymbol{z}_i^{*'} \boldsymbol{\kappa} + \varpi_i, \end{aligned} \tag{4}$$

with $\boldsymbol{x}_i^*$ and $\boldsymbol{z}_i^*$ vectors of individual covariates affecting Θ_i and η_i, respectively; $\boldsymbol{v}$ and $\boldsymbol{\kappa}$ regression coefficients, and ε_i and ϖ_i error components. Individual characteristics may be either directly observed or latent (i.e., measured by a set of observable indicators). In the latter case, a measurement specification is required, as is standard in the SEM framework. The resulting extension of the B-GLIRT model, incorporating latent regressions, is illustrated in Figure 3.

Latent variable interactions

In our model, we introduce item-specific interaction effects between item position and the latent traits (see

Eq. 1) to examine whether the relationship between item response and/or response time and its position is moderated by students' ability and/or speediness. Evaluating interaction effects is critical when assessing abilities using randomised item ordering, as highlighted by Bulut et al. (2017), particularly in applications like ours where item arrangement may influence response patterns.

From a modelling perspective, this entails estimating a latent interaction between a latent trait (ability or speediness) and an observed predictor (item position), treated as a single-indicator factor. Among the various methods available for adding interaction terms (see, for example, Aytürk et al., 2020; Klein & Moosbrugger, 2000; Marsh et al., 2004), we apply the product-indicator method based on the double mean centring approach (Lin et al., 2010; Little et al., 2006; Marsh et al., 2004). According to the product-indicator approach, we multiply the single-indicator item position by all the indicators of the latent moderator (ability or speediness). This yields product indicators that define the measurement model of the interaction factor. Specifically, we define J new latent variables $\Theta_i P_{ij}$ (with $j = 1, \dots, J$) to explore the moderation effect of ability, each measured by the set of indicators $(Y_{i1} \cdot P_{ij}, Y_{i2} \cdot P_{ij}, \dots, Y_{iJ} \cdot P_{ij})$. Similarly, we introduce as many J new latent variables $\eta_i P_{ij}$ with indicators $(T_{i1} \cdot P_{ij}, T_{i2} \cdot P_{ij}, \dots, T_{iJ} \cdot P_{ij})$ to investigate speediness moderation. Following the double mean-centring approach, all the first-order indicators are mean-centred before creating the product indicators that, in turn, are additionally mean-centred. This procedure does not impact the estimation of interaction effects. Still, it reduces model specification complexity by ensuring that first-order indicators (e.g., Y_{i1} and P_{i1}) are uncorrelated with the corresponding product indicators (i.e., $Y_{i1} \cdot P_{i1}$). In addition, equality constraints are applied to residual covariances between product indicators sharing a first-order indicator (e.g., $Y_{i1} \cdot P_{i1}$ and $Y_{i2} \cdot P_{i1}$; Schoemann & Jorgensen, 2021). To identify the scale of the latent variables and interaction effects, we also fix the factor loadings of the corresponding first indicator to 1 (Jöreskog & Yang, 2013).

Inference

In the proposed framework, parameter estimation may be performed using methods suitable for latent variable models, such as Maximum Likelihood Estimation (MLE; Wood et al., 2002), Diagonally Weighted Least Squares (DWLS; Savalei & Rosseel, 2014), or Bayesian approaches. Each method has its advantages and limitations. For instance, MLE is widely used due to its desirable asymptotic properties, but can become computationally intensive when dealing with high-dimensional latent variables. Alternatively, Bayesian estimation offers greater flexibility and the incorporation of prior information, but it requires careful consideration of priors and is computationally intensive. DWLS is particularly well-suited for analysing ordinal data, offering robust and efficient estimation under this setting. In this study, the estimation process employs the DWLS method to ensure accurate modelling.

To select the B-GLIRT model with the best fit, we may rely on standard measures used in the SEM setting, such as the Comparative Fit Index (CFI), the Tucker-Lewis index (TLI), and the Root Mean Square Error of Approximation (RMSEA). Relevant differences between models are usually indicated by changes in CFI greater than 0.020 and in RMSEA greater than 0.015 (Chen, 2007; Schermelleh-Engel et al., 2003; Vandenberg & Lance, 2000). Moreover, the χ^2 test—specifically, the robust χ^2 test (Satorra, 2000; Satorra & Bentler, 2001) when the DWLS method is used—can also be performed to compare nested models.

As the proposed models are embedded in the GLMM framework, they can be estimated using general statistical packages, such as `lavaan` (Rosseel, 2012) and `semTools` package (Jorgensen et al., 2021); this latter implements interaction effects with latent variables based on the product-indicator method above described. The script for implementing various specifications of the B-GLIRT model with `lavaan` is available at https://github.com/mariaiannario/MBR-Supplementary-Material-B-GLIRT.

Measurement invariance

Measurement invariance (MI) is a crucial statistical property that ensures the item parameters within a measurement model remain consistent, or invariant, across different groups or conditions (Mellenbergh, 1989; Meredith, 1993; Millsap & Yun-Tein, 2004). This property is particularly significant in psychological and social sciences, as it enables meaningful comparisons across populations (such as age or gender groups) based on item scores. Indeed, MI confirms that observed score differences reflect actual variations in the construct rather than measurement bias. Any differences in measurement model parameters between groups should be considered in the absence of MI.

The testing process for MI typically involves a hierarchical series of increasingly constrained models, where the baseline model is a fully unconstrained model with all item parameters differing across groups. Then, equality constraints are systematically applied to examine whether measurement parameters are equivalent across groups. These constraints can include equality of *(i)* factor loadings (metric invariance or weak factorial invariance), *(ii)* item intercepts (scalar invariance or strong factorial invariance), and *(iii)* residual variances (strict or residual invariance). The comparison in terms of fit between these nested models relies on the indices mentioned above, that is, differences in CFI, TLI, RMSEA, and changes in χ^2 values.

In the mainstream literature on factor analysis, the study of MI typically involves blocks of parameters (i.e., factor loadings, item intercepts, residual variances) across the entire set of items simultaneously. This approach yields a global evaluation of invariance, referred to as *full* MI. However, it does not allow for the identification of which individual items contribute to the lack of invariance, thereby limiting the possibility of establishing conclusions in terms of *partial* MI. This aspect has been largely overlooked in the context of factor analysis, with some notable exceptions (e.g., Byrne et al., 2013; Reise et al., 1993). In contrast, DIF analysis (Holland & Wainer, 1993) within the IRT setting explicitly addresses this issue, with item-level evaluation to detect whether specific items function differently across groups with respect to the latent traits and, hence, identify subsets of items that guarantee fairness in the measurement process. The process of testing for partial MI follows the same general principles as that of full MI described above. However, instead of evaluating parameter blocks for all items simultaneously, the analysis is performed sequentially, focusing on one item at a time. Typically, the baseline model used to evaluate DIF is a fully constrained model, where all item parameters are held equal across groups. A series of increasingly unconstrained models is then estimated, setting equal and unequal for each item, in turn, the factor loadings, the item intercepts, and the error variances (if present), while the parameters of the remaining elements (referred to as anchor elements) remain constrained.

It is worth noting that, in both full and partial MI analyses, the testing process follows a hierarchical structure: testing for equality of item intercepts (or residual variances) across groups is meaningful only if the factor loadings are first established as equal (Hancock et al., 2009; Millsap & Meredith, 2007; Thissen, 2025).

In the context of response accuracy and response times, testing for MI is critical to ensure that the assessment tool is fair and unbiased across different subpopulations. Although MI testing is standard for assessing latent abilities, its application in models that incorporate response times remains relatively rare (Glas & van der Linden, 2010; Molenaar et al., 2015). Models integrating item responses with response times present distinct challenges to achieving MI, as the inclusion of temporal information increases model complexity. Testing for MI in both the response accuracy and response time components is crucial to avoid biased group comparisons. If MI is violated for either element, it suggests that group differences may be influenced by factors other than the latent traits being measured, such as test-taking strategies or differential effects of test anxiety.

We recommend beginning the analysis by testing for full MI, relying on the measurement model in Eq. 1. A potential limitation of this approach is its reduced sensitivity to detect non-invariance when DIF affects only a small subset of items. In such situations, the presence of one or a few non-invariant items may be overshadowed by the large number of invariant items, leading global tests of MI to suggest an acceptable overall fit. As a consequence, meaningful sources of non-invariance may remain undetected. Therefore, reliance on full MI testing alone is not recommended; instead, partial MI solutions should also be explored. At this stage, there is no universally agreed strategy for how to proceed, as the appropriate approach depends on the specific objectives of the study. In principle, the procedure follows the hierarchical structure outlined above: testing item-by-item for equality of factor loadings, and, conditional on the loadings being invariant, for equality of item intercepts and residual variances. However, in the proposed B-GLIRT model, the presence of multiple parameter types (namely, two kinds of factor loadings (α_j and γ_j), a cross-relation parameter (ρ), two kinds of intercepts (τ_{jk} and λ_j), a residual response time variance (σ^2_{tj}), and up to four item position coefficients (ω_{1j}, ω_{2j}, ζ_{1j}, and ζ_{2j})) allows for multiple testing strategies, depending on the research objectives. In our case, the goal is to identify a subset of items that are unaffected by DIF, rather than to pinpoint the specific kinds of parameters where DIF occurs. Therefore, we adopt a blockwise testing approach within each item, beginning with the two loadings. If evidence of DIF is detected at this stage, the item is classified as non-invariant, and no further testing is conducted for that item. If no DIF is detected, we proceed to test the equality of the intercepts. Should DIF emerge at this stage, the procedure stops; otherwise, we continue by testing the equality of the residual variances in the response time submodel, and so on. It is worth noting that, although testing the equality of item position coefficients across groups is technically feasible, such comparisons may lack substantive interpretability, as discussed in the "Case Study" section. Particular attention should be given to the treatment of the cross-relation parameter between the two latent traits. The key question is whether this parameter should be considered part of metric invariance. If so, its equality should be tested alongside the factor loadings; if not, it may be tested separately after evaluating the other parameters. In either case, it is important to note that the cross-relation parameter is constant across items. Therefore, in the context of partial MI analysis, repeatedly testing its equality across groups is redundant and may lead to inconsistent or contradictory results. Given the lack of specific guidelines in the literature on this issue, we recommend conducting two parallel analyses: one assuming ρ is equal across groups, and one allowing ρ to differ.

Identifiability

Identifiability is a critical issue in these models. In the context of item response models, identifiability is often ensured by setting constraints on the latent variable distributions, such as fixing the mean and variance of the latent traits (e.g., assuming the ability and speed latent variables have a mean of zero and a variance of one). Additional constraints may be necessary for more complex models, such as those with interaction terms or hierarchical structures, to avoid parameter confounding. In B-GLIRT models, identifiability challenges arise due to the complexity introduced by jointly modelling item responses and response times, each influenced by underlying latent traits, typically ability and speed. Certain constraints must be imposed on the model parameters to address these challenges.

- *Latent Variable Constraints*: To ensure identifiability, it is common practice to standardise the distributions of the latent variables. Specifically, the ability (Θ_i) and speed (η_i) latent traits are assumed to follow a standard normal distribution, that is, $\Theta_i, \eta_i \sim N(0, 1)$. This scaling eliminates ambiguity in estimating the latent variable effects and ensures that the model parameters are on a comparable scale.
- *Cross-Relation Function*:
  The B-GLIRT model includes a cross-relation function, $f(\Theta_i, \rho)$, which connects the latent traits of ability and speed. To maintain identifiability, the form of this function must be carefully specified. Popular specifications include linear relations, such as the hierarchical model of van der Linden (2007) or the ability model of Thissen (1983). These predefined forms help constrain the model adequately, ensuring the parameters are uniquely estimable.
- *Item-Specific Effects*:
  Including item-specific effects, such as item difficulty and discrimination parameters, can complicate identifiability. To mitigate this, constraints are often applied. For example, item discrimination (α_j for item responses and γ_j for response times) can be equal between items or constrained within a specific range. In some cases, parameters may be fixed to particular items to anchor the scale of the model.
- *Interaction Terms*:
  Additional constraints are necessary when interaction effects between item position and latent traits (e.g., ability and speed) are included. For instance, item position effects may be assumed constant across items or set to zero if interactions are not significant. This reduces the risk of overparameterization, which can lead to non-identifiable models.

Additional constraints are applied at different steps to test for MI, as proposed by Molenaar et al. (2015). Particular attention is devoted to refining the procedure by incorporating recent developments from Svetina et al. (2020), specifically to address the challenges posed by ordinal observed variables.

In our B-GLIRT model, identifiability is achieved by imposing specific constraints on the cross-relational function $f(\theta_i; \rho)$ and on the parameters linking item responses, response times, and item positions, as detailed in the Case study section.

Case study

The case study considers students' ability assessment in Statistics through the Moodle digital platform, a widely used learning management system in higher education, which offers various configurations for learning assessments, including the ability to present questions to students randomly and to collect response times at the item level, in addition to item responses.

In what follows, we first describe the data used in the study and then illustrate and discuss the analysis results.

Sample and data

The study involves a convenience sample of $N = 308$ students in the psychology degree course at the University of Naples Federico II, attending the introductory Statistics course. Participants are predominantly female (71.3%) with age ranged between 18 and 47 (mean = 20, sd = 2.96). At the end of the learning module on descriptive statistics, students are asked to take a summary test to evaluate their acquired knowledge. Additionally, they are requested to complete a self-report questionnaire to gather socio-demographic information and explore psychological factors related to learning Statistics.

Students' ability in Statistics is measured using a multidimensional approach exploiting three of the five Dublin descriptors (Gudeva et al., 2012), which are general statements used to evaluate the knowledge depth a student has achieved within a specific topic at the end of each cycle of higher education studies. For each Dublin descriptor considered, a set of multiple-choice questions with four answer options has been designed; see Fabbricatore et al. (2023) and Fabbricatore et al. (2024) for examples of the questions. During the assessment, questions are presented to students randomly, and they are not permitted to revisit previous questions throughout the test. Students are allowed one hour to complete the test, autonomously deciding how much time to allocate to each question. Response coding is performed considering a partial credit for intermediate performance levels on items, moving beyond the binary right/wrong answer coding rationale. Accordingly, correct answers receive two credits, partially correct answers receive one credit, and incorrect answers receive no credit. Missing responses due to test closure for time constraints are treated as missing values.

The analysed data comprise students' responses to the 10 questions developed for the "applying knowledge and understanding" Dublin descriptor (see the Appendix for a full description of the questions), including item responses, response times, and item position information. This ability, which may be briefly named as *numeracy proficiency*, refers explicitly to applying knowledge to identify, analyse, and solve problems. Thus, our focus is mainly on students' numeracy and data interpretation skills. The attitude and affinity with numbers (numeracy) are essential to understanding the complexities of today's world, which is increasingly shaped by data-driven information. In this vein, studies show that numeracy proficiency is linked to better outcomes in various life domains, including financial literacy, health decisions, and career opportunities (Cordova Jr et al., 2024; Fabbricatore & Palazzo, 2023; Reyna et al., 2009; Skagerlund et al., 2018). However, its prevalence among youth remains limited, highlighting the need for further investigation to ensure students are adequately equipped to tackle future challenges. Specifically, our study aims to explore individual differences in response accuracy and response time, accounting for the potential influence of item position during assessment. Additionally, we are interested in examining the effect of individual characteristics on students' response behaviour and achievements, as the literature highlights their pivotal role. Among these factors, we consider gender to detect possible differences regarding measurement invariance of response accuracy and response times. Additionally, we investigate the impact of individual characteristics on students' numeracy proficiency and speediness. Among these, we focus on the role played by mathematics knowledge, self-efficacy, test anxiety, academic procrastination, number of days of study, and type of high school.

Regarding gender, studies have reported gender differences in Science, Technology, Engineering and Mathematics (STEM) performance (Cipora et al., 2015; Else-Quest et al., 2010), often favouring males, and have discussed the influence of gender stereotypes that frame math-related subjects as fields predominantly for men (Keller & Dauenheimer, 2003; Quinn & Spencer, 2001; Rodarte-Luna & Sherry, 2008).

Regarding psychological factors, self-efficacy positively influences students' academic performance by enhancing their enjoyment of learning, motivating them to apply statistical knowledge to real-world problems, and sustaining overall motivation to learn (Honicke & Broadbent, 2016; Richardson et al., 2012). Conversely, test anxiety and academic procrastination have detrimental effects. Test anxiety disrupts concentration and cognitive processing due to feelings of tension, worry, and over-arousal (Fabbricatore et al., 2023; Rajiah et al., 2014), while procrastination reduces the time available for mastering content by promoting a tendency to delay tasks (Steel et al., 2001). On the positive side, consistent study habits—such as studying regularly over several days—enhance achievement by reinforcing learning and promoting retention. Cognitive factors like prior mathematics knowledge are also crucial for success in Statistics, as research indicates that significant gaps in mathematical skills can create substantial barriers to understanding statistical concepts (Chiesi & Primi, 2010; Johnson & Kuennen, 2006; Rabin et al., 2018). Finally, the type of high school may also shape students' achievements, with students from scientific high schools reporting more substantial knowledge of STEM subjects than their peers from non-scientific high schools.

To investigate the effect of individual covariates, we consider the sub-sample of $n = 268$ students who completed the self-report questionnaire on psychological factors at the beginning of the course. To assess psychological factors, we use the following psychometric scales:

- the *Mathematical Prerequisites for Psychometrics* scale (PMP; Galli et al., 2008) to evaluate the basic mathematics abilities for an introductory Statistics course in a psychology degree program. The scale consists of 30 multiple-choice questions, and students gain one point for each correct response, whereas wrong and missing answers receive no credits. Thus, scores range between 0 and 30;
- the *Motivated Strategies for Learning Questionnaire* (MSLQ; Bonanomi et al., 2018) to measure students' self-efficacy and test anxiety. The Self-efficacy sub-scale consists of 9 items referring to expectancy for success in a knowledge domain, perceived ability to complete a task and self-confidence. The test anxiety sub-scale comprises 4 items that measure students' anxiety and concern about taking exams. Students are asked to indicate their degree of agreement on a 5-point Likert scale ranging from 1 (not at all true for me) to 5 (very accurate for me). Accordingly, the theoretical subscale range given by the sum of the scores observed on the corresponding items equals 9-45 for self-efficacy and 4-20 for test anxiety.
- the *Academic Procrastination Scale—Short Form* (APS-SF; Yockey, 2016; Fabbricatore et al., 2025) to assess academic procrastination, namely students' tendency to delay academic tasks intentionally. The scale consists of 5 items with a 5-point Likert response scale from 1 (I disagree) to 5 (I agree). The total score is defined as the arithmetic mean of the score observed on the single items (theoretical range 1-5), where higher scores indicate a greater tendency to delay academic tasks despite the adverse outcomes it causes (e.g., low academic performance, distress, anxiety).

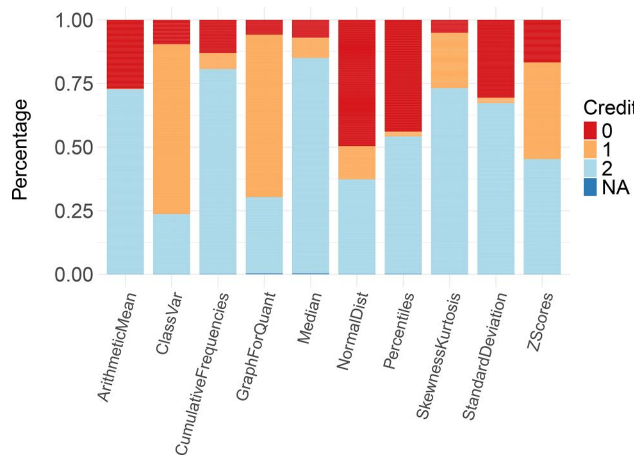

Figure 4. Response distribution. Note that no credit was given to wrong answers, one credit to partially correct answers, and two credits for totally correct answers. Missing responses due to time constraints were considered missing values (NA).

Descriptive statistics

Figure 4 depicts the response distribution for the 10 considered items assessing students' numeracy proficiency. Missing values refer to missing responses due to test closure for time constraints. Results show that arithmetic mean, cumulative frequencies, median, skewness and kurtosis are the most straightforward topics for students who report more than 70% of entirely correct answers for the corresponding items. Conversely, they experience more difficulties with normal distribution, z-scores, and percentiles.

Regarding item response times, Figure 5 shows the distribution for each item. From this figure, we observe that items requiring less time concern the classification of variables (average time = 82 seconds) and skewness and kurtosis (average time = 113 seconds). In contrast, those more time-demanding are percentiles (average time = 167 seconds) and z-scores (average time = 177 seconds).

Finally, descriptive statistics about cognitive and psychological factors related to learning Statistics are computed on the sub-sample of $n = 268$ students. Figure 6 shows the score distribution for the considered variables. According to the psychometric scale theoretical range, it can be seen that, on average, students report a medium-high level of mathematics knowledge with about 22 out of 30 correct answers, moderate self-efficacy and test anxiety, and medium-low academic procrastination levels. Hence, test

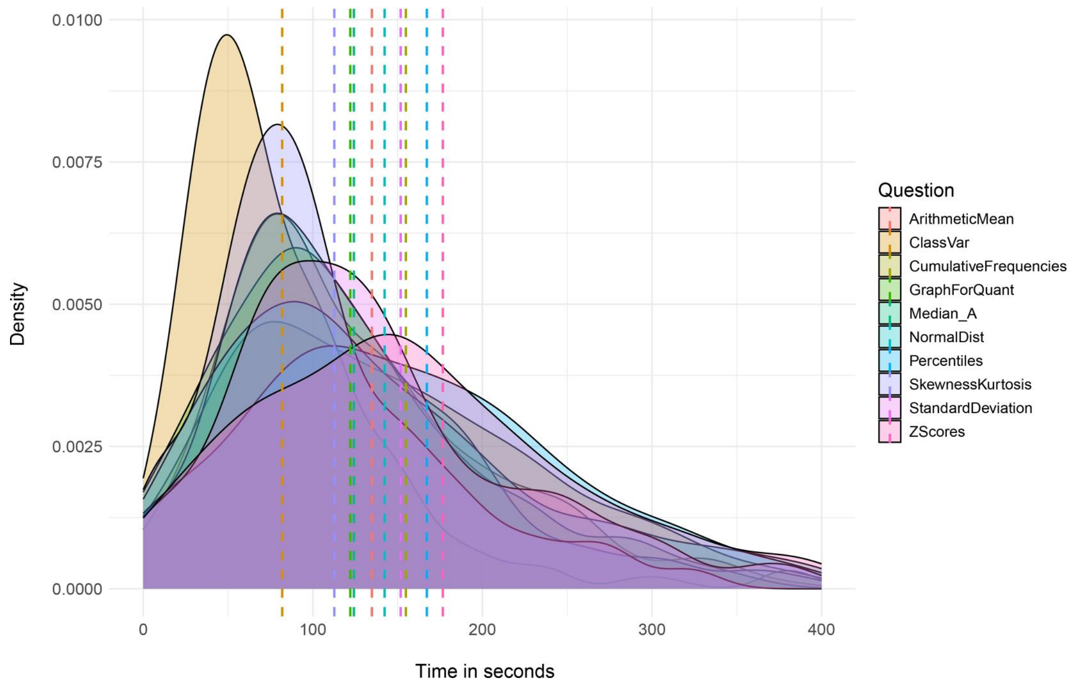

Figure 5. Response time distribution. Dashed lines indicate mean values.

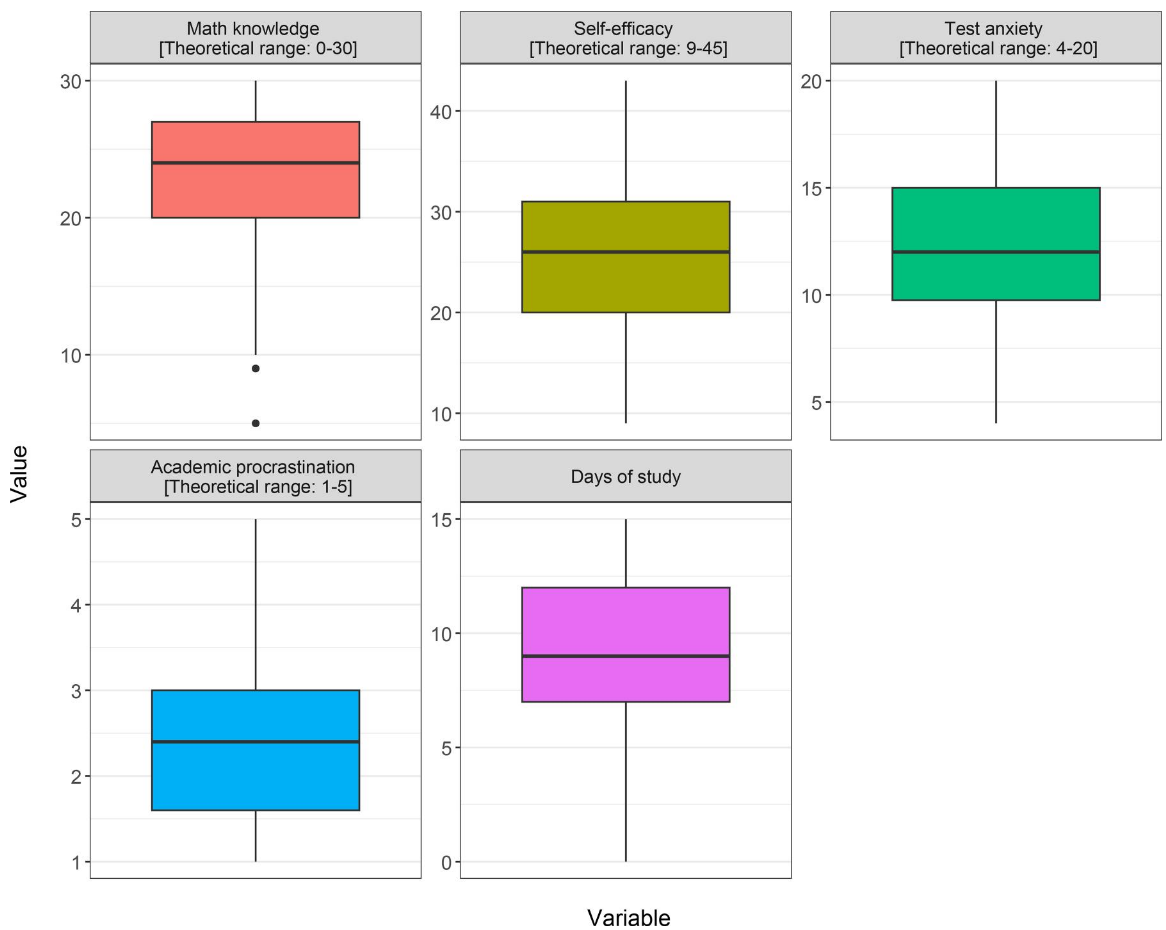

Figure 6. Boxplots depicting the distribution of the scores of the psychological variables and the number of study days. Note that the y-axis scales of the variables differ due to their different theoretical ranges.

anxiety generally has a low impact on students' emotions, and they maintain a relatively positive expectation of their ability to learn and succeed in Statistics. Moreover, on average, students report a low tendency to delay academic tasks intentionally and state that they had studied for nine days to prepare for the test.

B-GLIRT model selection

We estimate and compare several B-GLIRT models that incorporate item responses, response times, and item positions to identify the best model describing the data at hand. First, we select the best specification for the cross-relation function $f(\Theta_i; \rho)$ by comparing two popular B-GLIRT models that are the hierarchical model of van der Linden (2007), with $f(\Theta_i; \rho)$ as in eq. (2), and the ability model of Thissen (1983), with $f(\Theta_i; \rho)$ as in eq. (3). At this stage, we consider the less constrained model (eq. 1 without item positioning effect) with free discrimination parameters on both responses and response times. Model fit indices reported in Table 1 suggest the van der Linden specification is the best one, leading us to consider $f(\Theta_i; \rho) = -\gamma_j \rho \Theta_i$ in the following analyses.

Table 1. Model fit comparison for different specifications of the cross-relation function

	Df	CFI	TLI	RMSEA
van der Linden	349	0.904	0.948	0.023
Thissen	349	0.856	0.922	0.028

Once the best cross-relation function is selected, we determine the most suitable model parameter constraints. To this aim, we first compare a set of nested models (Mod0-Mod6) to investigate the position effect by keeping equal the discrimination parameters (i.e., $\alpha_j = \alpha$, $\gamma_j = \gamma$, for $j = 1, \dots, J$). Table 2 reports the details about the parameter setup of the considered B-GLIRT models.

Then, we compare the best model among Mod0-Mod6 with the same model having released discrimination parameters (Mod7), namely, free α_j and γ_j parameters for all j. To compare the models, we rely on differences in fit statistics (i.e., $\Delta\chi^2$, Δ CFI, $|\Delta$ RMSEA$|$) and the robust χ^2 test.

Results, reported in Table 3, suggest that Mod2 would be preferable among Mod0–Mod6. Although some less constrained models (Mod5 and Mod6) yield statistically significant improvements according to the chi-square difference test, the associated changes in model fit indices remain below conventional thresholds (i.e., 0.02 for Δ CFI and 0.015 for $|\Delta$ RMSEA$|$) for practical relevance. Therefore, to preserve model parsimony, Mod2 was retained as the reference model. Accordingly, Mod7 is defined starting from Mod2, releasing constraints on parameters α_j and γ_j. The last row of Table 3 shows that Mod7 must be preferred to Mod2 because it presents the best model fit according to the chi-square test and fit indices. Summarising, the best model includes a significant effect of position only on response times—where the position effect is

Table 2. Parameter setup for the position effect in the B-GLIRT models comparison. Note that for all models, the item discrimination parameters are constrained to be equal ($\alpha_j = \alpha$, $\gamma_j = \gamma$, for $j = 1, \dots, J$).

	Parameter setup		
Model	ω_{1j}	ζ_{1j}	Model description
Mod0	$\omega_{1j} = 0$	$\zeta_{1j} = 0$	There is no position effect.
Mod1	$\omega_{1j} = \omega_1$	$\zeta_{1j} = 0$	The position effect on responses is equal across items. There is no position effect on response times.
Mod2	$\omega_{1j} = 0$	$\zeta_{1j} = \zeta_1$	There is no position effect on responses. The position effect on response times is equal across items.
Mod3	$\omega_{1j} = \omega_1$	$\zeta_{1j} = \zeta_1$	The position effect on responses and response times is equal across items.
Mod4	$\omega_{1j} = \omega_{1j}$	$\zeta_{1j} = \zeta_1$	The position effect is item-specific for responses and equal across items for response times.
Mod5	$\omega_{1j} = \omega_1$	$\zeta_{1j} = \zeta_{1j}$	The position effect is equal across items for responses and item-specific for response times.
Mod6	$\omega_{1j} = \omega_{1j}$	$\zeta_{1j} = \zeta_{1j}$	The position effect is item-specific for responses and response times.

Table 3. Model fit comparison for nested B-GLIRT models (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

	Df	χ^2	CFI	TLI	RMSEA	Compared model	Δ CFI	$\|\Delta$ RMSEA$\|$	$\Delta\chi^2$
Mod0	387	976.31	0.017	0.518	0.070				
Mod1	386	974.84	0.018	0.517	0.070	Mod0	0.001	< 0.0001	0.28
Mod2	386	611.15	0.625	0.815	0.044	Mod0	0.608	0.026	122.33***
Mod3	385	609.67	0.625	0.815	0.044	Mod2	0.000	< 0.0001	0.65
Mod4	376	601.58	0.624	0.810	0.044	Mod2	0.001	0.005	7.84
Mod5	376	594.49	0.636	0.816	0.044	Mod2	0.011	< 0.0001	27.28**
Mod6	367	586.39	0.634	0.811	0.044	Mod2	0.009	< 0.0001	32.85*
Mod7	368	431.18	0.895	0.946	0.024	Mod2	0.270	0.020	387.70***

constrained to be equal across items—and free discrimination parameters.

Finally, we investigate the person-specific effect of item position on response times by adding an interaction term between position and speediness to the speediness measurement model of Mod7. The interaction effect is tested only on response times, as the main effect of item position on item responses was not supported. Moreover, the interaction term is imposed to be equal across items as emerged for the main effect. Results report a non-significant interaction effect between item position and speediness (estimated coefficient equals -0.003 with p-value $=$ 0.67), leading to the selection of Mod7 as the best model, that is,

$$\mathbb{E}\left[Y_{ij}^*\right] = \tau_{jk} + \alpha_j \Theta_i$$
$$\mathbb{E}\left[T_{ij}^*\right] = \lambda_j + \gamma_j \eta_i - \gamma_j \rho \Theta_i + \zeta_1 P_{ij}$$

In sum, after a careful model selection process, Mod7 is the best; the related parameter estimates are discussed below. Figure 7 reports the estimates of loading parameters α_j, γ_j, and $\gamma_j\rho$, while Figure 8 shows the item-step difficulty parameters τ_{jk}.

Figure 7 highlights that all the loading parameters are significant since confidence intervals do not contain 0. Moreover, the loadings of the item response indicators (blue solid circles) are generally higher than those of the corresponding response time indicators (blue empty circles). Finally, as expected, the response time indicators load more heavily on the speediness latent variable (yellow empty triangles) than numeracy proficiency (blue empty circles). The sign of the response time loadings for speediness is positive; as highlighted in the technical section, this indicates that the interpretation of the latent variable is reversed: lower scores indicate higher speediness, and vice-versa. The positive values of response time loadings, combined with the negative sign of the cross-relation parameter ($\hat{\rho} = -0.48$, p-value < 0.001), result in positive cross-loadings $-\gamma_j\rho$ (blue empty circles in the Figure 7). Note that, in the orthogonal SEM formulation we propose, cross-loadings should not be interpreted as direct effects. Rather, the direction of the association between numeracy proficiency and speediness is determined by the sign of ρ, given the scale of the speediness factor. Accordingly, higher levels of

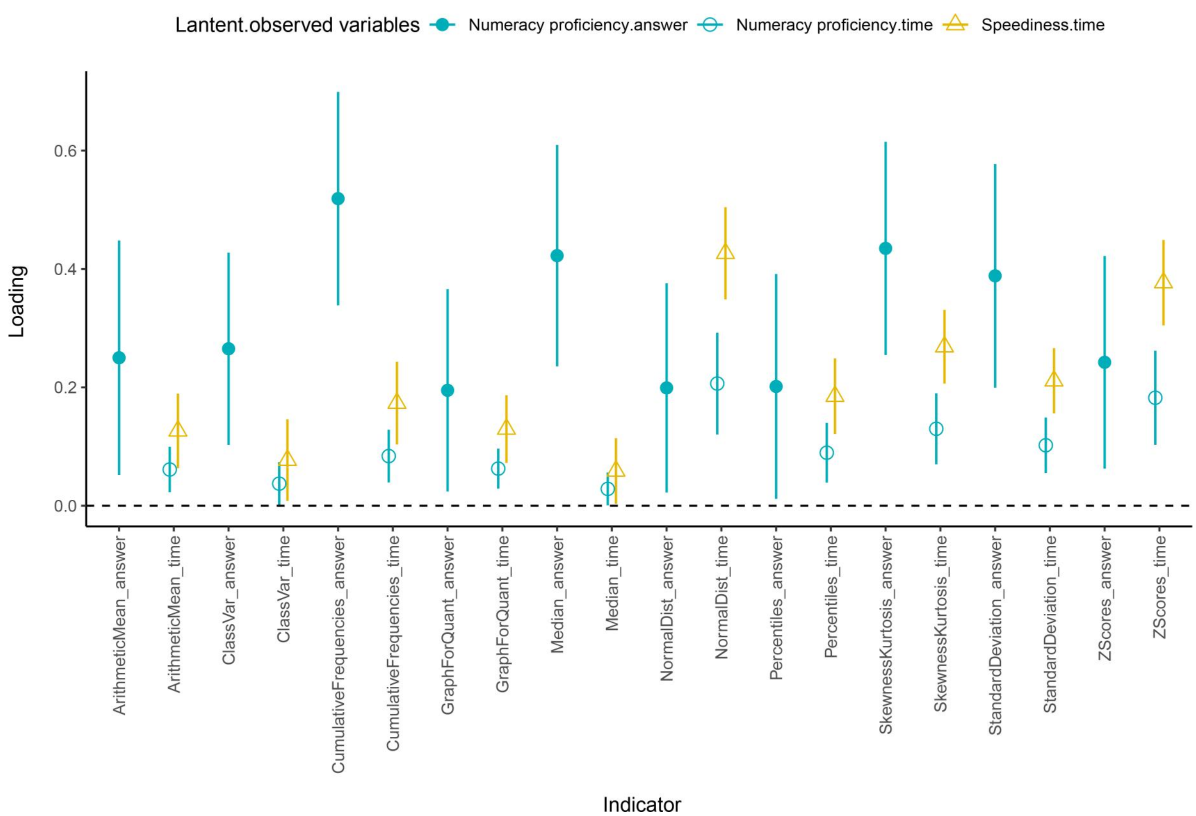

Figure 7. Loading parameter estimates (points) and 95% confidence intervals (lines): blue lines with blue solid circles for the loading of the item responses on numeracy proficiency (parameters α_j); yellow lines with yellow empty triangles for the loadings of response times on speediness (parameters γ_j); blue lines with blue empty circles for the loadings of responses times on numeracy proficiency (parameters $\gamma_j\rho$).

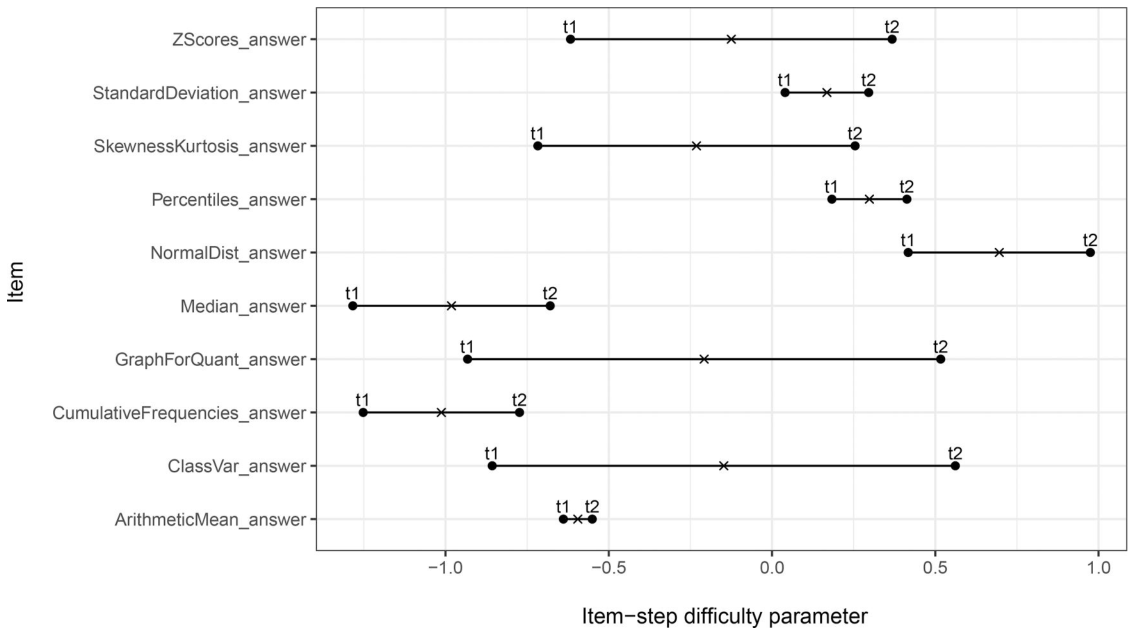

Figure 8. Estimates of item-step difficulty parameters τ_{jk}. The symbol $\times$ indicates item location (average item difficulty), whereas t1 and t2 denote the first and the second steps, respectively.

numeracy proficiency are significantly associated with higher speediness (i.e., low values of η_i).

Looking at Figure 8, where estimated item-step parameters τ_{jk} and their average values (item locations) are depicted, it is possible to notice that all items present ordered item-step difficulty parameters, supporting the accuracy of our response coding strategy. Moreover, item locations indicate that questions related to the median, cumulative frequencies, and arithmetic mean are the easiest. Conversely, students have greater difficulty dealing with topics such as standard deviation, percentiles, and normal distribution (most difficult items). The estimated response time intercepts $\hat{\lambda}_j$ range from 4.58 (ClassVar) to 5.42 (StandardDeviation).

Concerning the item position, the results show that it does not affect item responses but significantly impacts response time. Specifically, the regression coefficient is negative and equal across items ($\hat{\zeta}_1 = -0.03$, p-value < 0.001), suggesting that as the positions of the items increase, the response time decreases. Although the magnitude of the position effect may appear small, it aligns with the findings of Bulut et al. (2017), where the regression coefficients for the item position effect on item difficulty ranged from -0.009 to -0.037.

Covariate effects

To explore the impact of individual characteristics on students' latent traits, we introduce in the analysis a set of covariates. First, starting from our best model (Mod 7), we investigate the effect of gender on the measurement part of the model, checking for MI. Second, relying on the (sub)set of items for which (partial) MI holds, we consider the number of days of study, type of school, and all the psychological factors as predictors of the two latent variables (numeracy proficiency and speediness) in the structural part of the model. For these analyses, we use a subset ($n = 268$) of the whole sample of students who completed the questionnaire on psychological factors at the beginning of the course.

Covariate effects on indicators: MI by gender

Given the role of gender in shaping students' performance in STEM subjects, differences in favour of men regarding response accuracy and response times could also be expected in our case study. To determine whether these differences reflect bias in the measurement parameters of the model, we test for MI by gender, following the stepwise procedure described in the methodological section "Measurement invariance". We first assess whether full MI holds; we then examine the equality of measurement parameters between males and females, one item at a time. The parameters considered in this analysis include the cross-relation parameter, factor loadings, item intercepts and thresholds, and the residual variance of the time response submodel. The coefficient ζ_1, which represents the effect of item position on response time, is constrained to be equal across the two groups.

Table 4. Full MI analysis by gender: fit measures to compare hierarchically constrained models (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

	Restricted parameters.	Df	CFI	TLI	RMSEA	Compared model.	Δ CFI	|Δ RMSEA|	$\Delta\chi^2$
Submod7_1	–	737	0.753	0.873	0.046				
Submod7_2	α_j	746	0.739	0.867	0.047	Submod7_1	0.014	0.001	25.00*
Submod7_3	α_j, γ_j	755	0.669	0.833	0.053	Submod7_2	0.070	0.006	54.89***
Submod7_4	α_j, ρ	747	0.698	0.847	0.051	Submod7_2	0.041	0.004	8.11**

Table 5. Partial MI analysis by gender: items affected by DIF; fit criteria: Δ CFI, |Δ RMSEA|, $\Delta\chi^2$ (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

Unconstrained		Criteria				Estimates	
parameters	Item	Δ CFI	|Δ RMSEA|	$\Delta\chi^2$	Parameter	Male	Female
		Panel A): ρ *equal between males and females*					
α_j and γ_j							
	ZScores	0.019	0.002	3.75	α_j	−0.062	0.377***
					γ_j	0.165*	0.404***
	StandardDeviation	0.017	0.002	3.30	α_j	0.642**	0.289**
					γ_j	0.499***	0.184***
τ_{kj} and λ_j							
	NormalDist	0.011	0.001	3.60	τ_{j1}	2.554*	−0.905
					τ_{j2}	3.124**	−0.347
					λ_j	5.254***	4.954***
		Panel B): ρ *different between males and females*					
α_j and γ_j							
	GraphForQuant	0.024	0.002	3.90	α_j	0.183	0.169
					γ_j	0.419**	0.053
	ZScores	0.012	0.001	1.74	α_j	0.068	0.397***
					γ_j	0.173*	0.359***
τ_{kj} and λ_j							
	NormalDist	0.012	0.001	2.97	τ_{j1}	2.554*	−0.962
					τ_{j2}	3.124**	−0.404
					λ_j	5.386***	5.099***

Indeed, in our specific context, we find no compelling reason to assume that this effect differs between males and females. Nonetheless, there may be situations where testing the invariance of ζ_1 is warranted. For instance, in an educational setting where a test is administered to both students with specific learning disorders and those without, the effect of item position could plausibly vary between groups.

Table 4 displays the models' fit indices for the full MI analysis, whereas Table 5 refers to the results of the partial MI analysis. For the Full MI analysis we rely on the pre-specified ΔCFI threshold greater than 0.020. For the item-level checks, where freeing a single constraint typically yields smaller changes in global fit indices, we use a more sensitive screening threshold ($\Delta CFI > 0.010$; Putnick & Bornstein, 2016). In this way, we reduce the risk of overlooking single problematic items.

The full MI analysis (Table 4) begins by comparing the baseline model (Submod7_1), in which all parameters are freely estimated across the two groups, with model Submod7_2, which imposes equality constraints on the α_j loadings. Model Submod7_2 is then compared with the more restricted model Submod7_3, which additionally constrains the γ_j loadings to be equal, and with model Submod7_4, where both α_j and the cross-relation ρ parameter are constrained to be equal across groups. Based on the fit indices, there are indications of a lack of invariance in the loading parameters, other than in ρ. Specifically, changes in CFI (Δ CFI) and RMSEA (|Δ RMSEA|) suggest issues with the γ_j parameters and ρ, while the chi-square difference test (based on $\Delta\chi^2$) also points to potential problems with the α_j parameters. At this stage, the analysis proceeds item-by-item to identify a subset of items that satisfy invariance.

We proceed hierarchically by comparing a fully constrained model, in which all item parameters are held equal between males and females, with models in which parameters are freed one at a time, starting with the loadings. Following the recommendations outlined in the methodological section "Measurement Invariance", two parallel partial MI analyses are conducted: one assuming that ρ is equal across males and females (Table 5, Panel A), and the other allowing ρ to differ between the two groups (Table 5, Panel B). These analyses identify the following items as exhibiting a differential functioning: ZScores and NormalDist under both assumptions about ρ; additionally, StandardDeviation shows differential functioning when ρ is held constant, and GraphForQuant when ρ is allowed to vary. We note that the identification of items exhibiting DIF follows a conservative rule: an item is flagged as displaying DIF if at least one of the three criteria (Δ CFI, $|\Delta$ RMSEA$|$, $\Delta\chi^2$) indicates its presence.

The results in Table 5 indicate that the direction of the DIF is not systematically in favour of males or females throughout the test, but rather item-specific. This pattern suggests that the observed differential functioning may be related to factors such as previous learning experiences, learning styles, familiarity with the content, or preferred problem-solving strategies. For example, based on the item content reported in the Appendix, the differences observed for GraphForQuant may be linked to males' greater exposure to graph-related tasks in informal settings, as well as their generally higher spatial skills (Gallagher et al., 2000). Moreover, previous research has shown that males are more likely to adopt heuristic strategies in quantitative problem solving, whereas females tend to adhere more closely to classroom-learned procedures (Gallagher et al., 2000; Innabi & Dodeen, 2018; Zhu, 2007). In our case study, this gender-related tendency may contribute to the DIF observed for ZScores and NormalDist, which require multi-step reasoning typically learned during the course and less often experienced in daily-life settings, as well as for StandardDeviation, which is grounded in specific theoretical knowledge covered in the course (e.g., population vs. sample standard deviation formula). Conversely, items without evidence of DIF are related to more basic concepts and straightforward solution strategies that would seem to be shared by both genders.

In sum, the combination of the markedly different cross-relations between ability and speediness across groups, along with the lack of MI for certain items (primarily due to differing loading patterns between males and females), indicates that the bivariate construct defined by the two latent variables (ability and speediness) is substantively different across groups. The latent variables do not carry the same meaning individually, and should not be interpreted as equivalent. This between-group discrepancy exemplifies the "essential ambiguity" in the meaning of the latent variables in the bivariate analysis described by Thissen (1983). Consequently, the "effective ability" (Thissen, 1983) observed when respondents are allowed unlimited time may differ substantially from the construct traditionally referred to as "ability" in speeded achievement tests, such as the one employed in our study.

Based on the above considerations, we select as our final measurement model the subset of items that remain invariant when allowing ρ to vary between the two groups (see Table 5, Panel B). Parameter estimates are obtained using a multiple-group model in which all parameters are constrained to be equal across groups, except for ρ (Mod8). The model shows an acceptable overall fit: CFI $= 0.711$, TLI $= 0.863$, RMSEA $= 0.038$. Figure 9 reports the estimates of the loading parameters α_j and γ_j, while Figure 10 presents the estimates of item-step difficulty parameters τ_{jk}. The response time intercepts λ_j range from 4.62 (ClassVar) to 5.71 (StandardDeviation). The estimated parameter ρ is -0.03 (s.e. $= 0.12$; p-value $= 0.82$) for males and -0.41 (s.e. $= 0.15$; p-value $= 0.006$) for females, suggesting that individual differences in response times are negatively correlated to numeracy proficiency in the female group, whereas they are uncorrelated in the male group. The item position effect on response time is estimated at $\zeta_1 = -0.024$ (s.e. $= 0.002$; p-value < 0.001).

To conclude this section, we note that the proposed MI procedure is general and can, in principle, be applied to other grouping variables. In the present study, we illustrate its use, focusing on gender, which is theoretically and empirically central to STEM performance. As an additional robustness check, we also replicated the MI analysis using test anxiety as a grouping variable and found no evidence of non-invariance. The corresponding results are available from the authors upon request.

Covariate effects on latent variables

To investigate the effect of individual characteristics on the two considered latent variables (numeracy proficiency and speediness), we enter in Mod8 a set of covariates as predictors of the latent variables, as in eq. (4). Specifically, we consider the type of high school, mathematics knowledge, self-efficacy in Statistics, test anxiety, academic procrastination, and the number of days of study; for the psychological factors, we model the corresponding items explicitly, rather than relying on summarised scores, as shown in Figure 3. Table 6 shows the estimates of regression coefficients, along with standard errors, z-statistics and p-values for the selected model containing only covariates whose estimated coefficients are significant at 5%. Note that, as with all parameters except ρ, the covariate effect on the latent variables is also assumed to be equal across groups.

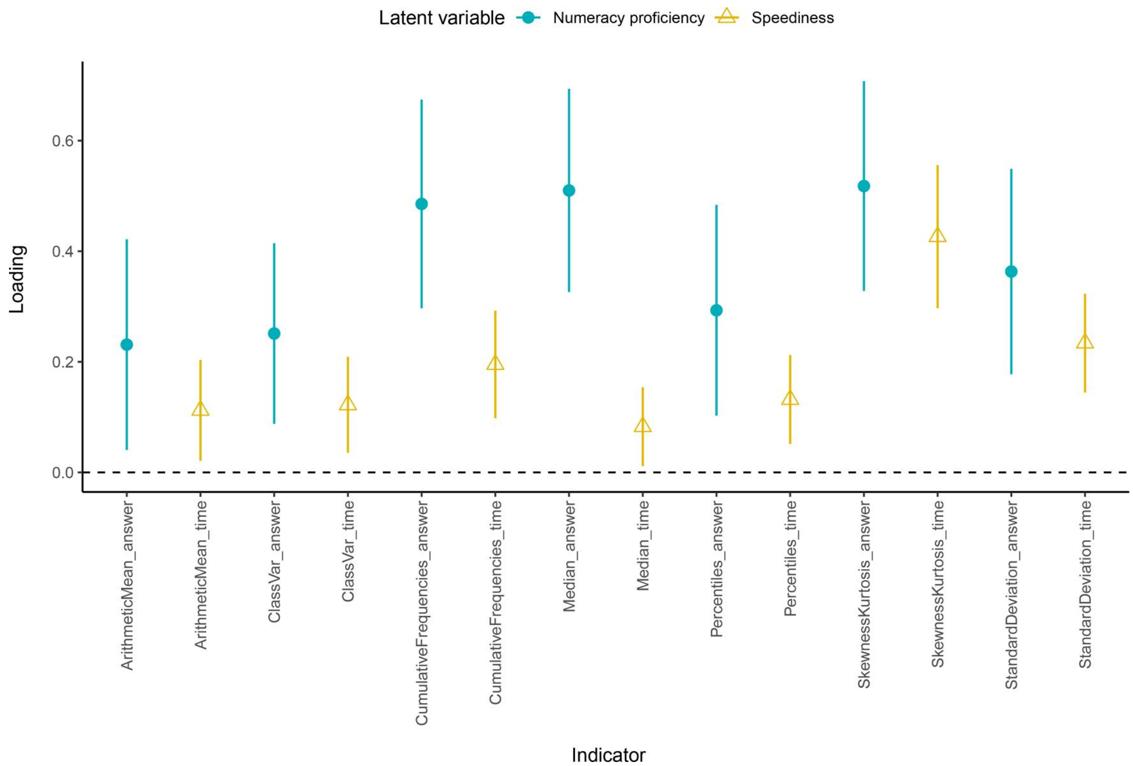

Figure 9. Loading parameter estimates (points) and 95% confidence intervals (lines) of the multiple-group model with selected items (Mod8): blue lines with blue solid circles for the loading of the item responses on numeracy proficiency (parameters α_j); yellow lines with yellow empty triangles for the loadings of response times on speediness (parameters γ_j).

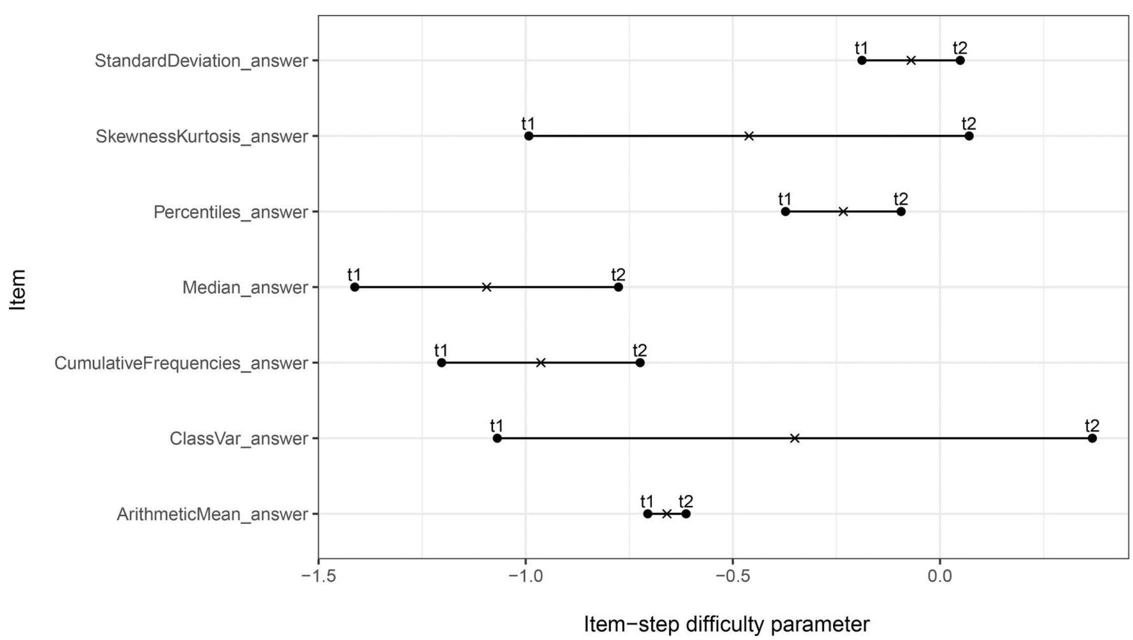

Figure 10. Estimates of item-step difficulty parameters τ_{jk} for the multi-group model with selected items (Mod8). The symbol $\times$ indicates item location (average item difficulty), whereas t1 and t2 denote the first and the second steps, respectively.

Mathematics knowledge and attending a scientific high school show the strongest positive association with numeracy proficiency ($\hat{\nu}_1 = 0.734$, $p = 0.001$; $\hat{\nu}_2 = 0.806$, $p = 0.020$). These results align with numerous studies that emphasise the foundational role of prior mathematical knowledge in developing further ability in STEM subjects (Chiesi & Primi, 2009; Lavidas et al., 2020), and with the more

Table 6. Covariates effect on latent variables: regression coefficients (Coeff.), standard errors (s.e.), z-statistics (z-stat), and *p*-values.

Latent variable	Covariate	Coeff.	s.e.	z-stat	*p*-value
Numeracy proficiency	Math knowledge	0.734	0.231	3.178	0.001
	Scientific high school	0.806	0.347	2.321	0.020
	Self-efficacy	0.384	0.118	3.249	0.001
	Days of study	0.150	0.050	3.011	0.003
	Academic procrastination	−0.355	0.127	−2.798	0.005
Speed	Scientific high school	−0.829	0.404	−2.052	0.040
	Test anxiety	0.423	0.181	2.330	0.020

intensive STEM curriculum typically characterising scientific high schools, particularly in mathematics, physics, and the natural sciences. Belief in one's capabilities and expectancy for success (self-efficacy) also has a positive effect on numeracy proficiency ($\hat{\nu}_3 =$ 0.384, p-value $=$ 0.001), in line with the social cognitive theory (Bandura, 1997) and previous literature (Elbulok-Charcape et al., 2024; Hoegler & Nelson, 2018; Zare et al., 2011). In addition, a greater number of study days is positively associated with numeracy proficiency ($\hat{\nu}_4 = 0.150$, p-value $= 0.003$), highlighting the beneficial effect of consistent study habits on the development of numeracy skills. Conversely, academic procrastination significantly negatively affects numeracy proficiency ($\hat{\nu}_5$ = -0.355, p-value = 0.005). This finding is consistent with other studies (Kim & Seo, 2015; Steel, 2007), highlighting the detrimental effects of procrastination on performance. Regarding speediness, among all the considered covariates, only the type of high school attended by the students and test anxiety had a significant effect. Specifically, results highlight that students previously enrolled in scientific high school tend to take less time to complete the test (regression coefficient for speediness: $\hat{\kappa}_1$ = -0.829, p-value $=$ 0.040; note that the interpretation is positive, namely, scientific high school students have a higher speediness). These students may be more accustomed to math-related tasks and calculations, which likely enhances their speed in processing similar information and contributes to faster problem-solving. Conversely, higher levels of test anxiety are associated with lower speediness ($\hat{\kappa}_2 = 0.423$, p-value $= 0.020$; the positive sign reflects the reversed meaning of η_i), in line with prior research showing that anxiety disrupts concentration and reduces the efficiency of the central executive component of working memory (Derakshan & Eysenck, 2009).

Conclusions

This study presents an innovative framework for jointly modelling response accuracy, response times, and item position effects in computer-based learning assessments. By employing a Bivariate Generalised Linear Item Response Theory (B-GLIRT) model, we have provided a comprehensive approach to simultaneously account for numeracy proficiency and speed, offering insights into the intricate dynamics of test-taking behaviour. The proposal is based on latent bivariate IRT modelling for responses-and-time in a Structural Equation Modeling (SEM) framework.

The proposal contributes in three ways to the literature on item position effects in large-scale assessments. First, it allows researchers and practitioners to examine both linear position effects and interaction effects in the same model. Second, the approach presented in this study is a flexible method for studying position effects with various item response models for dichotomously and polytomously scored items—such as the two-parameter model, Partial Credit Model, and Graded Response Model, among others. Regarding this last point, the study also offers, for the first time, an ordinal scale-based approach for assessing the accuracy of responses. Third, the approach makes it possible to assess whether the item position on difficulty or response time gives important signals on the need to ensure a fair distribution in tests when randomised among respondents. Importantly, our contribution lies in methodologically integrating item position and response time into a latent variable modelling framework, rather than treating them as separate or sequential phenomena. While previous studies have acknowledged the relationship between position and speededness, our model offers a structured, inferential approach to assess their joint effects within the GLIRT class. In our specific context, the analysis reveals that while item position does not significantly affect item difficulty, it affects response times, with shorter times taken for items appearing later in the test. This decrease reflects the influence of time pressure or test fatigue. Conversely, the lack of a significant effect of item position on response accuracy suggests that varying the order of test items in the proposed assessment is not an issue, supporting this practice to enhance test security, as it does not influence final scores. We acknowledge that response times are not routinely used for test scoring; however, incorporating them into the modelling framework provides a more refined understanding of performance patterns and test-taking behaviour. This is particularly relevant in large-scale assessments, where position-related fatigue or disengagement could bias proficiency estimates. Our results emphasise the necessity of incorporating item position and response times into the modelling framework to understand assessment performance better.

We also discussed measurement invariance (MI) for the proposed model, with particular attention to the treatment of the cross-relation function between the two latent variables. Specifically, we outline a strategy based on testing for Full MI and Partial MI, that is, invariance at the item level. The item-level analysis makes it possible to identify subsets of items that are invariant across groups, which can be used to obtain comparable measures of the latent traits.

Furthermore, including individual and item-specific covariates, such as mathematical knowledge, self-efficacy, and psychological factors, highlights the complex interplay between cognitive, motivational, and behavioural influences on academic outcomes. The reduced response times as item positions advance reflect the growing concern about time as the test progresses. Our approach provides a statistically rigorous tool for quantifying and modelling these phenomena, thereby improving the accuracy and fairness of inferences drawn from test scores. It encourages the development of more adaptive and equitable testing strategies by offering practical guidance on the effects of item order and the relevance of individual differences.

The findings and limitations of this study point to several directions for future research. While the proposed framework accounts for item position and response time effects, future work could extend the model by incorporating additional predictors of test performance, such as cognitive load, linguistic complexity, and task-specific characteristics (e.g., different weights attributed to each item). Including these factors may help clarify why certain items are particularly sensitive to position effects.

Further research is also needed to assess the applicability of the model across different testing formats, including computer-based adaptive testing, in which item selection depends on the test-taker's ongoing performance. Investigating how position effects interact with adaptive algorithms could provide valuable insights for optimising item selection and enhancing assessment fairness.

Finally, assessing the generalisability of the proposed approach across different educational contexts, such as high-stakes versus low-stakes standardised testing, would provide further evidence on its practical value and would support the development of more adaptive and equitable assessment practices. More generally, the proposed framework is not limited to educational measurement. Its joint modelling of latent ability, response speed, and sequential effects makes it potentially relevant for a wide range of applications in which performance accuracy and response processes jointly reflect underlying individual differences. Such contexts include psychological diagnostics, vocational assessment, personnel selection, and clinical or behavioural testing. To demonstrate this broader scope, the Supplementary Material presents an additional application of the B-GLIRT model outside the educational domain. Specifically, the model is applied in a behavioural research setting examining individuals' ability to attribute sex to visual stimuli, thereby illustrating the flexibility and wider applicability of the proposed modelling framework.

Article information

Conflict of interest disclosures: No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: Silvia Bacci acknowledges the financial support by the Italian Ministry of University and Research (MUR) under grant Department of Excellence project 2023-2027 ReDS "Rethinking Data Science" – Department of Statistics, Computer Science, Applications – University of Florence (CUP B13C220004500001).

Rosa Fabbricatore acknowledges the financial support provided by the European Union -NextGenerationEU, in the framework of the GRINS - Growing Resilient, Inclusive and Sustainable project (GRINS PE00000018 - CUP E63C22002140007).

Maria Iannario gratefully acknowledges the support of the Marie Sklodowska-Curie Actions under the European Unions Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance (acronym: DIGITAL, project no. 101119635).

Silvia Bacci and Maria Iannario acknowledge the support of PRIN 2022 call. Project 'Latent variable models and dimensionality reduction methods for complex data' (PI Prof. Paolo Giordani), Project No. 20224CRB9E, CUP E53C24002270006.

Role of the funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: Authors acknowledge that the questions whose responses are analysed in the manuscript were developed as part of the 'Adaptive LEArning in Statistics' (ALEAS) Erasmus + project (KA+2018-1-IT02-KA203-048519). The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the funding agencies is not intended and should not be inferred. The authors would like to thank Stefano Federici (University of Perugia, IT) for making the data analysed in the Supplementary Material available for use.

ORCID

Silvia Bacci http://orcid.org/0000-0001-8097-3870
Rosa Fabbricatore http://orcid.org/0000-0002-4056-4375
Maria Iannario http://orcid.org/0000-0002-2646-9937

References

Adams, R. J., Wilson, M. R., and Wang, W. C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement*, 21:1–24. https://doi.org/10.1177/0146621697211001

Albano, A. D. (2013). Multilevel modelling of item position effects. *Journal of Educational Measurement*, 50:408–426. https://doi.org/10.1111/jedm.12026

Alexandrowicz, R. and Matschinger, H. (2008). Estimation of item location effects by means of the generalized logistic regression model: a simulation study and an application. *Psychological Science Quarterly*, 50:64–74.

Aytürk, E., Cham, H., Jennings, P. A., and Brown, J. L. (2020). Latent variable interactions with ordered-categorical indicators: Comparisons of unconstrained product indicator and latent moderated structural equations approaches. *Educational and Psychological Measurement*, 80(2):262–292. https://doi.org/10.1177/0013164419865017

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Becker, B., van Rijn, P., Molenaar, D., and Debeer, D. (2022). Item order and speededness: implications for test fairness in higher educational high-stakes testing. *Assessment and Evaluation in Higher Education*, 47:1030–1042. https://doi.org/10.1080/02602938.2021.1991273

Bollen, K., Rabe-Hesketh, S., and Skrondal, A. (2008). Structural equation models. In J. M. Box-Steffensmeier, H. Brady, and D. Collier (Ed.), *Oxford handbook of political methodology* (pp. 432–455). Oxford University Press.

Bonanomi, A., Olivari, M. G., Mascheroni, E., Gatti, E., and Confalonieri, E. (2018). Using a multidimensional Rasch analysis to evaluate the psychometric properties of the motivated strategies for learning questionnaire (MSLQ) among high school students. *Testing, Psychometrics, Methodology in Applied Psychology*, 25(1):83–100.

Bulut, O., Quo, Q., and Gierl, M. J. (2017). A structural equation modelling approach for examining position effects in large-scale assessments. *Large-scale Assessments in Education*, 5:1–20. https://doi.org/10.1186/s40536-017-0042-x

Bungaro, L., Desimoni, M., Matteucci, M., et al. (2024). Jointly exploring mathematics ability and speed in large-scale computer-based testing. *Statistical Methods and Applications*, 33:1429–1450. https://doi.org/10.1007/s10260-024-00762-0

Byrne, B. M., Shavelson, R. J., and Muthén, B. (2013). Testing for the equivalenceof factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105:456–466. https://doi.org/10.1037/0033-2909.105.3.456

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. https://doi.org/10.1080/10705510701301834

Chiesi, F. and Primi, C. (2009). Assessing statistics attitudes among college students: Psychometric properties of the Italian version of the survey of attitudes toward statistics (sats). *Learning and Individual differences*, 19(2):309–313. https://doi.org/10.1016/j.lindif.2008.10.008

Chiesi, F. and Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal*, 9(1):6–26. https://doi.org/10.52041/serj.v9i1.385

Cipora, K., Szczygieł, M., Willmes, K., and Nuerk, H.-C. (2015). Math anxiety assessment with the abbreviated math anxiety scale: Applicability and usefulness: Insights from the polish adaptation. *Frontiers in psychology*, 6: 1833. https://doi.org/10.3389/fpsyg.2015.01833

Clayton, D. (1996). Generalized linear mixed models. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo methods in practice* (pp. 275–303). Chapman & Hall.

Cordova Jr, N., Kilag, O. K., Andrin, G., Tañiza, F. N., Groenewald, E., and Abella, J. (2024). Leadership strategies for numeracy development in educational settings. *Excellencia: International Multi-disciplinary Journal of Education (2994-9521)*, 2(1):58–68.

Davey, T. and Lee, Y.-H. (2011). Potential impact of context effects on the scoring and equating of multistage GRE revised general test. Technical report, ETS Research Report Series, ETS RR-11-26. https://doi.org/10.1002/j.2333-8504.2011.tb02262.x

De Boeck, P. and Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10:102. https://doi.org/10.3389/fpsyg.2019.00102

De Boeck, P. and Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, Code Snippets*, 48(1):1–28. https://doi.org/10.18637/jss.v048.c01

De Boeck, P. and Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

Debeer, D. and Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational*

Measurement, 50:164–185. https://doi.org/10.1111/jedm.12009

Derakshan, N. and Eysenck, M. W. (2009). Anxiety, processing efficiency, and cognitive performance: New developments from attentional control theory. *European psychologist*, 14(2):168–176. https://doi.org/10.1027/1016-9040.14.2.168

Duncan, O. (1975). *Introduction to structural equation models*. Academic Press.

Elbulok-Charcape, M., Grandoit, E., Krishnan, A., Bergdoll, R., Fogel, J., and Rabin, L. (2024). Changes in undergraduate students' self-efficacy and outcome expectancy in an introductory statistics course. *Graduate Student Journal of Psychology*, 22. https://doi.org/10.52214/gsjp.v22i1.12178

Else-Quest, N. M., Hyde, J. S., and Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological bulletin*, 136(1): 103. https://doi.org/10.1037/a0018053

Fabbricatore, R., Di Mari, R., Bakk, Z., de Rooij, M., and Palumbo, F. (2024). Students' proficiency evaluation: A non-parametric multilevel latent variable model approach. *Studies in Higher Education*, 1–28. https://doi.org/10.1080/03075079.2024.2386623

Fabbricatore, R. and Palazzo, L. (2023). Multilevel IRT models to explore heterogeneity in subjective financial knowledge at individual and regional levels: the Italian case. *Frontiers in Applied Mathematics and Statistics*, 9: 1278146. https://doi.org/10.3389/fams.2023.1278146

Fabbricatore, R., Parola, A., Pepicelli, G., and Palumbo, F. (2023). Understanding undergraduate statistics performance through a PLS-PM approach: the role of math knowledge, test anxiety and attitudes toward Statistics. *Statistica Applicata–Italian Journal of Applied Statistics*, 34(2):1–30.

Fabbricatore, R., Parola, A., Pepicelli, G., and Palumbo, F. (2025). Academic procrastination scale: Psychometric evidence for the Italian version based on CFA and IRT models. *Italian Journal of Applied Statistics*. https://doi.org/10.36253/ijas-17022

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., and Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of experimental child psychology*, 75(3):165–190. https://doi.org/10.1006/jecp.1999.2532

Galli, S., Chiesi, F., and Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the Rasch Model. *Testing, Psychometrics, Methodology in Applied Psychology*, 15(1):1–16.

Glas, C. A. and 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. https://doi.org/10.1348/000711009X481360

Goldberger, A. (1972). Structural equation models in the social sciences. *Econometrica: Journal of Econometric Society*, 40:979–1001. https://doi.org/10.2307/1913851

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and 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):608–626. https://doi.org/10.1037/a0034716

Gudeva, L. K., Dimova, V., Daskalovska, N., and Trajkova, F. (2012). Designing descriptors of learning outcomes for higher education qualification. *Procedia-Social and Behavioral Sciences*, 46:1306–1311. https://doi.org/10.1016/j.sbspro.2012.05.292

Hancock, G. R., Stapleton, L. M., and Arnold-Berkovits, I. (2009). The tenuousness of invariance tests within multisample covariance and mean structure models. In T. Teo, and M. S. x Khine (Eds.), *Structural equation modeling in educational research: Concepts and applications* (pp. 137–174). Sense Publishers.

Hartig, J. and Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, 54:418–431.

Hoegler, S. and Nelson, M. (2018). The influence of anxiety and self-efficacy on statistics performance: A path analysis. *Psi Chi Journal of Psychological Research*, 23(5). https://doi.org/10.24839/2325-7342.JN23.5.364

Holland, P. W. and Wainer, H. (1993). *Differential item functioning*. Erlbaum, Hillsdale, NJ.

Honicke, T. and Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational research review*, 17:63–84. https://doi.org/10.1016/j.edurev.2015.11.002

Innabi, H. and Dodeen, H. (2018). Gender differences in mathematics achievement in Jordan: A differential item functioning analysis of the 2015 timss. *School Science and Mathematics*, 118(3-4):127–137. https://doi.org/10.1111/ssm.12269

Johnson, M. and Kuennen, E. (2006). Basic math skills and performance in an introductory statistics course. *Journal of statistics education*, 14(2). https://doi.org/10.1080/10691898.2006.11910581

Jöreskog, K. G. and Yang, F. (2013). Nonlinear structural equation models: The Kenny–Judd model with interaction effects. In *Advanced structural equation modeling* (pp. 57–88). Psychology Press.

Jorgensen, T., Pornprasertmanit, S., Schoemann, A., and Rosseel, Y. (2021). semTools: Useful tools for structural equation modeling [R package semTools version 0.5-4]. Comprehensive *R* Archive Network *(CRAN)*.

Keller, J. and Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*, 29(3):371–381. https://doi.org/10.1177/0146167202250218

Kim, K. R. and Seo, E. H. (2015). The relationship between procrastination and academic performance: A meta-analysis. *Personality and individual differences*, 82:26–33. https://doi.org/10.1016/j.paid.2015.02.038

Klein, A. and Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the lms method. *Psychometrika*, 65(4):457–474. https://doi.org/10.1007/BF02296338

Lavidas, K., Barkatsas, T., Manesis, D., and Gialamas, V. (2020). A structural equation model investigating the impact of tertiary students' attitudes toward statistics, perceived competence at mathematics, and engagement on statistics performance. *Statistics Education Research Journal*, 19(2): 27–41. https://doi.org/10.52041/serj.v19i2.108

Leary, L. F. and Dorans, N. J. (1985). Implications for altering the context in which test items appear: a historical perspective on an immediate concern. *Review of Educational Research*, 55:387–413. https://doi.org/10.2307/1170392

Li, F., Cohen, A., and Shen, L. (2012). Investigating the effecr of item position in computer-based tests. *Journal of Educational Measurement*, 49:362–379.

Lin, G.-C., Wen, Z., Marsh, H. W., and Lin, H.-S. (2010). Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies. *Structural Equation Modeling*, 17(3):374–391. https://doi.org/10.1080/10705511.2010.488999

Little, T. D., Bovaird, J. A., and Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural equation modeling*, 13(4):497–519. https://doi.org/10.1207/s15328007sem1304_1

Marsh, H. W., Wen, Z., and Hau, K.-T. (2004). Structural equation models of latent interactions: evaluation of alternative estimation strategies and indicator construction. *Psychological methods*, 9(3):275. https://doi.org/10.1037/1082-989X.9.3.275

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, linear, and mixed models*. Wiley.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13(2):127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bullettin*, 115:300–307. https://doi.org/10.1037/0033-2909.115.2.300

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58:525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. and Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck and R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 131–152). Lawrence Erlbaum Associates Publishers.

Millsap, R. E. and Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3):479–515. https://doi.org/10.1207/S15327906MBR3903_4

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80:993–997. https://doi.org/10.1080/01621459.1985.10478215

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, 11:81–91.

Molenaar, D., Tuerlinckx, F., and van der Maas, 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

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65:391–411. https://doi.org/10.1007/BF02296153

Nagy, G., Nagengast, B., Becker, M., Rose, N., and Frey, A. (2018). Item position effects in a reading comprehension test: an IRT study of individual differences and individual covariates. *Psychological Test and Assessment Modeling*, 60:165–187.

Naumann, J. and Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, 53:1–16. https://doi.org/10.1016/j.lindif.2016.10.002

Ong, T. Q. and Pastor, D. A. (2022). Uncovering the complexity of item position effects in a low-stakes testing context. *Applied Psychological Measurement*, 46:571–588. https://doi.org/10.1177/01466216221108134

Pomplun, M. and Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *Journal of Educational Computing Research*, 30:243–254. https://doi.org/10.2190/Y4FU-45V7-74UN-HW4T

Putnick, D. L. and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41:71–90. https://doi.org/10.1016/j.dr.2016.06.004

Quinn, D. M. and Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues*, 57(1):55–71. https://doi.org/10.1111/0022-4537.00201

Rabin, L., Fink, L., Krishnan, A., Fogel, J., Berman, L., and Bergdoll, R. (2018). A measure of basic math skills for use with undergraduate statistics students: The macs. *Statistics Education Research Journal*, 17(2):179–195. https://doi.org/10.52041/serj.v17i2.165

Rajiah, K., Coumaravelou, S., and Ying, O. W. (2014). Relationship of test anxiety, psychological distress and academic motivation among first year undergraduate pharmacy students. *International Journal of Applied Psychology*, 4(2):68–72.

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, 78:538–544. https://doi.org/10.1007/s11336-013-9324-6

Ranger, J. and Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, 71(2):389–406. https://doi.org/10.1177/0013164410382895

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage, Newbury Park, CA.

Reckase, M. (2009). *Multidimensional item response theory*. Springer.

Reise, S. P., Widaman, K. F., and Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114:552–566. https://doi.org/10.1037/0033-2909.114.3.552

Reyna, V. F., Nelson, W. L., Han, P. K., and Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological bulletin*, 135(6):943. https://doi.org/10.1037/a0017327

Richardson, M., Abraham, C., and Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological bulletin*, 138(2):353. https://doi.org/10.1037/a0026838

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8:185–205. https://doi.org/10.1037/1082-989x.8.2.185

Rodarte-Luna, B. and Sherry, A. (2008). Sex differences in the relation between statistics anxiety and cognitive/learning strategies. *Contemporary Educational Psychology*, 33(2): 327–344. https://doi.org/10.1016/j.cedpsych.2007.03.002

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2): 1–36. https://doi.org/10.18637/jss.v048.i02

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). Springer. https://doi.org/10.1007/978-1-4615-4603-0_17

Satorra, A. and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4):507–514. https://doi.org/10.1007/BF02296192

Savalei, V. and Rosseel, Y. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1):1–14. https://doi.org/10.1080/10705511.2013.824793

Schermelleh-Engel, K., Moosbrugger, H., Müller, H., et al. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, 8(2):23–74.

Schoemann, A. M. and Jorgensen, T. D. (2021). Testing and interpreting latent variable interactions using the semtools package. *Psych*, 3(3):322–335. https://doi.org/10.3390/psych3030024

Schweizer, K. and Ren, X. (2013). The position effect in tests with a time limit: the consideration of interruption and working speed. *Psychological Test and Assessment Modeling*, 55:62–78.

Skagerlund, K., Lind, T., Strömbäck, C., Tinghög, G., and Västfjäll, D. (2018). Financial literacy and the role of numeracy–how individuals' attitude and affinity with numbers influence financial literacy. *Journal of behavioral and experimental economics*, 74:18–25. https://doi.org/10.1016/j.socec.2018.03.004

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models*. Chapman and Hall.

Steel, P. (2007). The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological bulletin*, 133(1):65. https://doi.org/10.1037/0033-2909.133.1.65

Steel, P., Brothen, T., and Wambach, C. (2001). Procrastination and personality, performance, and mood. *Personality and individual differences*, 30(1):95–106. https://doi.org/10.1016/S0191-8869(00)00013-1

Svetina, D., Rutkowski, L., and Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using MPLUS and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1):111–130. https://doi.org/10.1080/10705511.2019.1602776

Thissen, D. (1983). Timed testing: an approach using item response testing. In D. J. Weiss (Ed.), *New horizons in testing: latent trait theory and computerized adaptive testing* (pp. 179–203). New York Academic Press.

Thissen, D. (2025). A review of some of the history of factorial invariance and differential item functioning. *Multivariate Behavioral Research*, 60:211–235. https://doi.org/10.1080/00273171.2024.2396148

Trendtel, M. and Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009-2015 data. *Psychological Test and Assessment Modeling*, 60:241–263.

Trendtel, M. and Robitzsch, A. (2021). A Bayesian item response model for examining item position effects in complex survey data. *Journal of Educational and Behavioral Statistics*, 46:34–57. https://doi.org/10.3102/1076998620931016

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., Klein Entink, R. H., and Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34(5):327–347. https://doi.org/10.1177/0146621609349800

Vandenberg, R. J. and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1): 4–70. https://doi.org/10.1177/109442810031002

Vida, L. J., Bolsinova, M., and Brinkhuis, M. J. S. (2021). Speeding up without loss of accuracy: item position effects on performance in university exams. In *Proceedings of the 14th international conference on Educational Data Mining (EDM 2021)* (pp. 454–460). International Educational Data Mining Society.

Wang, L. (2019). Does rearranging multiple choice response options affect item and test performance. Technical report, ETS Research Report Series, ETS RR-19-02. https://doi.org/10.1002/ets2.12238

Weirich, S., Hecht, M., and Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, 38:535–548. https://doi.org/10.1177/0146621614534955

Wood, R., Wilson, D. T., Gibbons, R. D., Schilling, S. G., Muraki, E., and Bock, R. D. (2002). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Scientific Software International, Inc.

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-Scale Assessment in Education*, 7:1–21.

Yockey, R. D. (2016). Validation of the short form of the academic procrastination scale. *Psychological reports*, 118(1): 171–179. https://doi.org/10.1177/00332941 15626825

Zare, H., Rastegar, A., and Hosseini, S. M. D. (2011). The relation among achievement goals and academic achievement in statistics: the mediating role of statistics anxiety and statistics self-efficacy. *Procedia-Social and Behavioral Sciences*, 30:1166–1172. https://doi.org/10.1016/j.sbspro.2011.10.227

Zhu, Z. (2007). Gender differences in mathematical problem solving patterns: A review of literature. *International Education Journal*, 8(2):187–203.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictor. *Psychometrika*, 56(4):589–600. https://doi.org/10.1007/BF02294492

Appendix

In the following, we present the 10 questions developed to evaluate students' proficiency in the "applying knowledge and understanding" descriptor. Response option *a*) denotes the fully correct answer, while option *b*) indicates a partially correct response. Items exhibiting differential item functioning, whether ρ is equal or unequal between groups, are marked with an asterisk (*).

Arithmetic Mean

Three groups of students, with sizes 72, 44, and 52, have an average level of education (in years) of 9, 15, and 17, respectively. Which of the following is the best information about the average level of education of the 168 students?

a. It is equal to 13.05
b. It is equal to 15
c. It is equal to 13.67
d. It is equal to 9

ClassVar

The Italian Statistical Society is conducting a study on quality of life in 2022, which includes various dimensions such as economic conditions, health, education, and the environment. Question number 23 in the Economic Conditions section of the survey is: *"Please indicate your gross annual income in 2021: - - - -€"*. Which type of variable best represents this measure?

a. Ratio scale
b. Quantitative in general
c. Discrete
d. Interval scale

Cumulative Frequencies

A teacher asked a group of students how many hours they had studied the day before an exam. The results are summarized in the following table:

Number of hours	Cumulative absolute frequency
1–2	12
3–4	20
5–6	45
7–8	60
More than 8	78

What is the relative frequency of students who reported studying 5–6 hours?

a. 0.32
b. 0.58
c. 0.6
d. 0.45

GraphForQuant*

The owner of a restaurant wants to evaluate the effectiveness of the service they offer to their customers. For this reason, he decides to represent, using a chart, the waiting time (in minutes) that customers wait to be served on a typical day, which is reported in the following table:

Waiting time (in minutes)	Number of customers
0–4.59	9
5–9.59	7
10–14.59	25
15–19.59	18
20–24.59	16
25–29.59	5

What is wrong with this graphical representation?

a. It is not possible to use a bar chart to represent a continuous quantitative variable
b. It is not wrong to use a bar chart to represent a continuous quantitative variable, but a histogram would be more appropriate
c. The percentages do not add up to 100%
d. It is not appropriate to group the waiting time into intervals

Median

Customers waiting in line at a bank have complained about the amount of time they have to wait. The bank managers, beginning to investigate the problem, recorded a sample of customer waiting times at the bank. Here are the waiting times (in minutes):

26, 16, 26, 23, 15, 27, 20, 25

Which of the following is the most accurate information regarding the median of this dataset?

a. It is equal to 24
b. It is equal to 19
c. It is equal to 4
d. It is equal to 22.25

Normal Dist*

The average hourly wage of workers in a fast-food restaurant is €6.6, and we assume it is normally distributed with a standard deviation of $\sigma = 1.07$ euros. A worker earning €5.5 corresponds to which percentile of the distribution?

a. Equal to the 15th percentile
b. Equal to the 85th percentile
c. Equal to the 1.03rd percentile
d. Equal to the 10.7th percentile

Percentiles

A group of 40 students was asked how many hours they usually sleep each night. The results are shown in the frequency table below:

Hours	Frequencies
5	3
6	5
7	7
8	12
9	11
10	2

Which of the following statements best describes the first quartile of the reported hours of sleep per night?

a. it is equal to 7
b. it is less than 9
c. it is equal to 2
d. it is equal to 5

Skewness Kurtosis

A variable has a mean of $\overline{x} = 22$, a median of $Me = 18$, and a variance of $\sigma^2 = 252.53$. The value of the Hotelling–Solomon skewness coefficient will be equal to:

a. 0.252
b. 0.016
c. 4
d. −4

StandardDeviation*

A mathematics teacher wants to calculate the standard deviation of the grades obtained by his 20 students on the end-of-year exam. Considering that the students represent the entire population of interest, and knowing that the sum of the squared deviations from the mean is equal to 96.55, what will be the solution obtained by the mathematics teacher?

a. The standard deviation is equal to 2.197
b. The standard deviation is equal to 2.254
c. The standard deviation is equal to 4.827
d. The standard deviation is equal to 5.082

ZScores*

Intelligence test scores are normally distributed with a mean of 108 and a variance of 14. To compare these scores with those from another test, the following transformation is applied: $Y = 2 + 3X$, where X is the original variable.

Which of the following statements about the mean and variance of Y is correct?

a. The mean is equal to 326 and the variance is equal to 126
b. The mean is equal to 326 and the variance is equal to 130
c. The mean is equal to 108 and the variance is equal to 14
d. The mean is equal to 108 and the variance is equal to 130