Statistical Methods & Applications
https://doi.org/10.1007/s10260-024-00762-0

ORIGINAL PAPER

Jointly exploring mathematics ability and speed in large-scale computer-based testing

Luca Bungaro[1] · Marta Desimoni[2] · Mariagiulia Matteucci[1] · Stefania Mignani[1]

Accepted: 20 July 2024

Abstract

In large-scale tests, the implementation of computer-based testing (CBT) allows to automatically collect data not only on the students' response accuracy (RA) based on item responses of the test, but also on their response time (RT). RTs can provide a more comprehensive view of a test-taker's performance beyond just what is obtainable based on correct responses alone. In this paper a joint approach is considered to improve the estimation of ability scores involving complex data coming from computer-based test administration. The study focuses on analysing the data of Italian grade 10 mathematics national assessment administered by the National Institute for the Evaluation of the Education and Training System (INVALSI). In addition, a bivariate multilevel regression with speed and ability estimates, obtained by joint model, is developed including individual covariates to evaluate the contribution of individual and contextual variables in predicting test-taking speed and ability. Overall, the main results indicate that mathematics ability and speed are significantly and negatively correlated, and that the hierarchical data structure (students nested into classes) should be taken into account when explaining the dependency of ability and speed on explanatory variables, such as prior achievement, test anxiety, sociodemographic covariates, class compositional variables, school tracks and geographical area.

Keywords Computer-based testing · Response times · Multilevel models · Standardized mathematics tests

✉ Mariagiulia Matteucci
m.matteucci@unibo.it

[1] Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, 40126 Bologna, Italy

[2] National Institute for the Evaluation of the Education and Training Educational System (INVALSI), Rome, Italy

Published online: 30 July 2024

1 Introduction

International and national large-scale assessments (LSAs) have been utilized for many years to track students' knowledge and abilities. These evaluations are commonly known as low-stakes assessments, as their objective is to produce group-level scores for various populations and test performance has low, if any, personal consequences for test-takers.

Ability evaluation is a complex process requiring an adequate test tool (test) and the knowledge of individual and contextual characteristics to make more accurate inferences about ability. Considering additional data sources, such as test-takers' sociodemographic characteristics (e.g., gender, immigration status, socio-economic background), emotional-attitudinal characteristics (e.g., test anxiety), as well as data on test-taking behaviours can be useful to study the heterogeneity of outcomes and interpret them for use in educational policies. In addition, data from educational LSAs often have a hierarchical structure (student, class, school…) which makes specific methodological approaches necessary.

In recent years, the implementation of digital-based assessment (DBA) has been receiving a growing interest because of its operational advantages. DBA allows to automatically collect data not only on the students' response accuracy (RA) based on item responses of the test, but also on their response time (RT). The fundamental idea in response modelling is that a higher level of ability is associated to a higher probability of providing a correct response. Similarly, this concept can be applied to response time modelling where higher speed of working is linked to a lower expected response time.

Using the RTs, the assessment results based on item responses can be further improved in terms of precision and fairness, and the calibration costs could be reduced (van der Linden et al. 2010). RT data could indeed be useful to address several issues: verify the quality of the items; define test length and time limit; analyse the pacing and the strategy of test-takers; increase the efficiency of test scoring; detect test security (for example, cheating) and identify motivational behaviours. RTs can also reveal new information about test characteristics, test-takers' response behaviour, and test-takers' ability that would not be identified when using response information only. Overall, RTs can provide a more comprehensive view of a test-taker's performance beyond just what is obtainable based on correct responses alone (van der Linden et al. 2010; Bolsinova and Tijmstra 2018).

The aim of this study is to propose the jointly estimation of latent ability and latent speed from RA and RT data of the national assessment carried out by the National Institute for the Evaluation of the Education and Training System (INVALSI) in Italy, and to investigate the covariates of both ability and speed while taking into account the hierarchical structure of the data. In particular, the study will focus on data collected on students attending the second year of upper secondary school (Grade 10) in the school year 2017–2018. The main research questions focus on both methodological aspects and empirical issues.

In the first step of the analysis, following the approach of van der Linden (2007) and Fox and Marianti (2016), the ability and the speed are estimated jointly within a Bayesian hierarchical modelling framework. At level 1, the RTs and RA are assumed to be conditionally independently distributed given the speed and ability parameters. At level 2, the covariance between the speed and ability parameters, as well as a covariance matrix for the item parameters, are specified. The two covariance structures introduce a relationship between the RA and the RT data. In this way, the relation between speed and accuracy is considered during the estimation process and the accuracy of item and person parameters is improved. By applying the hierarchical model to data from the entire item set of mathematics items of the INVALSI design, the study aims to contribute to the stream of literature on the between-person speed-ability association in the mathematics domain and under low-stakes conditions.

In the second step of the analysis, given the hierarchical nature of data (students nested in classes), the ability and speed estimates are included as dependent variables in a bivariate multilevel model to evaluate the contribution of individual and contextual variables in predicting test-taking speed and ability. The speed measured from RTs is treated as a fundamental part of the test-taker's performance and, unlike what is usually examined in the literature, the interest is on investigating covariates of both test-taking speed and ability, also moving beyond the individual characteristics of the respondents. As observed by Costa and Chen (2023), literature on covariates of latent variables or other indices computed from RTs and other process data is still at its infancy. Furthermore, although recent ecological model suggests the adoption of multilevel models to investigate process data (e.g., Reis Costa and Leoncio Netto 2022), the contextual predictors of test-taking speed are still relatively unexplored. Therefore, with this study, we also aim to contribute to this stream of research by investigating the unique contribution of individual characteristics and contextual variables in predicting test-taking speed and ability in the mathematics domain under low-stakes conditions. In particular, we will focus on students' test anxiety, past achievement experiences (i.e., grade retention and the final mark obtained at the First-cycle State Leaving Examination) and sociodemographic characteristics (i.e., gender, socio-economic and cultural background, immigrant background). We also explored the role of the classroom composition, while taking into account the school track and the geographical area of the school.

The paper is organized as follows. Section 2 discusses the role and the importance of RTs in LSAs and contextualizes the current study in relation to the relevant literature. Section 3 describes the data used in the analysis. The methods used in the two-step analysis are introduced in Sect. 4. Section 5 reports the results and the discussion while Sect. 6 concludes the paper with limitations and further research for the current proposal.

2 Response time and test-taking speed in large-scale assessment scenario

LSAs typically require respondents to answer to power tests, with the time limits chosen so that each test-taker gets opportunity to attempt all the items in his or her test-form. The main focus of the power-test is on the given answers to the assessment item (namely, product data). Even when test takers are materially limited in time, time is not considered a central aspect of the test. The transition LSA from paper and pencil to DBA has allowed process data to be extracted from computer-generated logfiles. Among this data, RT has received an increased interest. There are several ways to define and calculate RT variables. For example, in the 2018 OECD's Programme for International Student Assessment (PISA), multiple RT variables are available, such as time spent on the last (or first) visit to an item and RT to first action, and RT across all item visits. In the IEA's TIMSS 2019 process data, the total time spent by the student on each item screen is provided. In the INVALSI national DBA, students can visit items multiple times, and, to date, the item RT variable corresponds to the sum of all the multiple time periods spent on the item.

Among advantages of having parallel data on responses and RTs in LSA is the possibility to measure both the ability and the speed of the test-takers. LSA design typically requires students to answer to a different subset of items. Therefore, measuring speed is not straightforward. For instance, from only raw scores obtained from RTs, we would not distinguish a student responding slowly from a subset of items being more time consuming. The joint hierarchical modelling of responses and RTs (van der Linden 2007; Fox et al. 2007; van der Linden and Fox 2016) is considered one of the most relevant approaches to disentangle item and person characteristics, thus estimating person latent speed and ability (controlling for item effects). Therefore, its application to process and product data from LSA may help to gather information on test-takers individual differences in underlying ability and working speed. Further, it allows to explore the nature of the between-person speed–ability association, along with correlates of these latent dimensions.

2.1 Between-person speed-ability association

In experimental psychology, a within-person Speed–Accuracy Trade-off (SAT) has been widely documented. In some situations, a test-taker may increase his/her RT at the cost of reducing the responses accuracy, while in other situations he/she responds more slowly to make relatively fewer errors (Zimmerman 2011). Van der Linden (2007) argued that the SAT is a within-person phenomenon. When switching from the within-person to the between-person level, the correlation between test-takers effective working speed on the test and their effective ability can be positive, negative, or even close to zero (van der Linden 2009). For instance, some recent studies found that under low-stakes the faster test-takers were also those with lower ability estimate in problems solving (Goldhammer et al. 2014), collaborative problem solving (De Boeck and Scalise 2019) and mathematics (Costa and Chen 2023; Zhang et al. 2024) domains, however an opposite association emerged on reading literacy (Goldhammer et al. 2014).

An explanation of these contrasting findings is based on a dual processing theory (Goldhammer et al. 2014; Naumann and Goldhammer 2017). A positive association between test-takers working speed and their performance might emerge when the task domain allows for automatic processing, so that students with higher ability are also those who can automatize the information processes, thus working faster on test. An example here is reading ability, that in skilled reader is well automatized. Otherwise, when information processing elements are less amenable to automatic processing and the task requires deliberate, controlled cognitive processing (Schneider and Shiffrin 1977; Shiffrin & Schneider 1977), or metacognitive processing (Pressley et al. 1989; Winne and Hadwin 1998), a negative association is likely to emerge between speed and ability, so that test-takers who spend more time on items are more likely to be those with better performance and higher (effective) ability. For instance, this might be the case of mathematics and problem-solving tests.

2.2 Covariates of test-taking speed under low-stakes condition

Individual differences in test-taking behaviour could arise from different psychological and cognitive factors (Reis Costa and Leoncio Netto 2022). We will focus on students' test anxiety, previous achievement-related experience, sociodemographic characteristics as covariates of their working speed on the test.

Previous findings have consistently reported a negative and statistically significant association between test anxiety and performance outcomes in standardized tests (see the meta-analysis by von der Embse et al. 2018). These results mostly refer to test accuracy scores and students' ability. Test anxiety could be also linked to working speed on test through cognitive and/or motivational mechanisms. For instance, the Processing Efficiency Theory (PET, Eysenck and Calvo 1992) and the Attentional Control Theory (ACT, Eysenck et al. 2007) predict that test-anxiety reduces cognitive efficiency, particularly on complex tasks. Anxiety-related thoughts burden working memory capacity, thus anxious test-takers plausibly need significantly more effort and response time to maintain performance effectiveness. Further test anxiety creates attentional bias for off-task threatening stimuli, making the ability to inhibit attention to distraction and shift between tasks more laborious and time consuming.

The association between test anxiety and test taking behaviour could be also examined under the test-taking motivation research tradition. In the Expectancy-Value Theory (EVT, Eccles and Wigfield 2002, 2020; Wigfield and Eccles 2000), test anxiety is conceptualized as an emotional cost of the test taking situation. Cost, importance, interest, utility are aspects of the value component of the EVT that, together with expectancy, are supposed to impact test-taking effort. Under low-stakes conditions, anxious individuals might wish to finish as soon as possible a test, because the uncomfortable testing situation cost too much emotionally relative to its benefits. Consistently, Akhtar and Firdiyanti (2023) found that test anxiety negatively predicts response-time effort in low-stakes conditions, so that high anxious test-takers tend to work faster on items. Noteworthy, some studies (Cassady and Finch 2020; Jerrim 2022) suggested to not limit the study of the association between test anxiety and other variables to linear relationships. Accordingly, the association between test anxiety and test-taking speed may depend upon the amount of anxiety experienced.

Besides test anxiety, also test-takers' past achievement-related experience are plausible associated with their test-taking behaviour. Students prior experience of academic success or failure, including both objective information (e.g. grades, retention) and subjective evaluations of achievements, are likely to shape their academic self-concept (Eccles et al. 1983; Wigfield and Eccles 2000; Bandura 1997; Skaalvik 1997). Therefore, those struggling academically might have lower expectations to perform well on the test, thus they may decide to put in less effort, especially under low-stakes conditions. This aligns with the EVT (Eccles and Wigfield 2002, 2020). Low effort might result in faster work on items, at the expense of accuracy, particularly when the task requires controlled and effortful processing. Extant research on predictors of test-taking behaviour has often neglected the role of students past academic experience, therefore its role in predicting test-taking speeds warrants more research.

Students' working speed on the test might be also associated with their sociodemographic characteristics (Costa and Chen 2023). For instance, the EVT highlights the role of students' family background in shaping their expectation of success and subjective task value (Eccles and Wigfield 2020). In line with this prediction, Teig et al. (2020) found that students with lower socio-economic and cultural background (ESCS) were more likely to belong to the disengaged latent profile, characterized for the low time-on-task, when taking PISA scientific inquiry tasks. Costa and Chen (2023) explored the covariates of test taking latent speed, ability and exploration behaviour using product and process data of a small subset of mathematics items from PISA 2012. The authors found that students' ESCS was the only sociodemographic variable to account for students' differences in students' test-taking speed. However, this effect was significant only in one of the three countries. In the same study, the authors explored the role of immigrant background, however no differences emerged between students with immigrant background and the natives in their test-taking speed.

Recent findings provided evidence of a different testing behaviours across gender groups under low-stakes conditions (Balart and Oosterveen 2019; Azzolini et al. 2019; Teig et al. 2020; Rutkowski et al. 2023; Marcq and Braeken 2023). These results have been ascribed to gender differences in several factors, such as work-avoidance, test-taking styles, and personality traits (DeMars et al. 2013). Girls were also found to have an advantage on boys on planning abilities (Naglieri and Rojahn 2001) self-regulatory behaviour, and time-management skills, which in turn could lead to gender differences in test-taking behaviour (Balart and Oosterveen 2019). However, previous results on gender differences in test-taking speed are not consistent across studies. Some previous studies reported that females work slowly on the test (Kroehne et al. 2019; Kapoor et al. 2023), while other studies found no significant differences across gender groups (e.g., Costa and Chen 2023).

Besides students' individual characteristics, it could be argued that test-taking speed could also be influenced by contextual factors. For instance, in their ecological framework for the analysis of process data, Reis Costa and Leoncio Netto (2022) advanced that variability in test-taking behaviours can be explained by variables laying at different layers, such as items and test characteristics, test-takers personal characteristics, classroom/school contexts, family/other outside of school ecology, educational system/state/country context. Extensive research has shown that students' performance in achievement tests are related to the peer composition of the learning environment (see the meta-analysis by van Ewijk & Sleegers (2010). For instance, peer-ESCS was found to affect students' cognitive and noncognitive outcomes, such as academic self-concept, over and above the effects of their own ESCS (Hansen et al. 2022). Further, as suggested by Ketonen and Hotulainen (2019), classmates can provide information about what is valued and important in relation to particular tasks and activities. Therefore, as predicted by the EVT (Eccles and Wigfield 2002, 2020), peers might affect students' willingness to strive in low-stakes testing situations. To date, the role played by contextual variables in shaping test-taking speed is relatively unexplored. Extending the study of the covariates of test-taking speed to characteristics of class/school (and above) might allow to depict a more comprehensive picture of students' performance in low-stakes assessments.

3 Data: the INVALSI mathematics test

INVALSI every year assesses all students attending grade 2, 5, 8, 10 and 13 in Italy. The main purpose of the INVALSI national assessment is to monitor the quality of the Italian educational system. From the school year (SY) 2017–2018, the INVALSI national testing program for students attending secondary school has undergone a main change: the transition to DBA. In this study, we use the mathematics data for grade 10 administered at the end of 2017–2018 SY (INVALSI 2018b). The theoretical framework (INVALSI 2018a) of the INVALSI mathematics assessment is aligned with the Italian National Guidelines for the curriculum. The assessment results are reported as proficiency score and in terms of descriptive achievement levels. The INVALSI DBA at the end of grade 10 is low-stakes at respondent level since it does not have consequences for the test-takers.

In the INVALSI test the number of involved examinees is very large and tests must be administered in multiple sessions and locations. Moreover, INVALSI need to produce several test forms to overcome security concerns, such as cheating and leaking of information. For grade 10, multiple test forms with prespecified characteristics are assembled from an item bank through automated test assembly. The test forms are assembled to fulfil measurement targets with respect to the expected test difficulty level and the precision on the ability continuum. Also, specific constraints on the test design are considered, such as the item type, the item domain, the item exposure, and the fulfilment of linking procedures. The grade 10 mathematics item bank consists of 143 items which are binary scored. The item difficulty parameters are estimated according to the Rasch model (Rasch, 1960). The difficulty estimates range from -2.431 to 2.907, with mean equal to 0.048 ($sd=0.941$) (Desimoni 2019).

The tests are administered to the whole student population, around 500,000 students. INVALSI also builds a random sample of around 41,000 units. The sampling procedure is a two-stage with Italian geographical region and school track stratification at the first stage. The units of the first stage are the schools and the units of the second stage are the classes. In this paper we analyse the results of the sample. Noteworthy, the INVALSI computer-based tests are conceptualized as power tests, not as speed tests. INVALSI imposes a time limit of 90 min on grade 10 tests, which is considered enough for students to read and answer all the questions.[1]

4 The methodological tools

4.1 Joint model for responses and response times

Response time modelling approaches can be classified into possibly overlapping and not necessarily homogeneous categories (De Boeck and Jeon 2019; van der Linden 2009): the response time models considering only the RT as outcome variable; the joint models including other variables (e.g. accuracy); dependency models in which RTs and other data are jointly modelled with the possibility of dependency beyond captured by latent variables; RTs as covariate models where RT is the independent variable and accuracy as outcome variable (De Boeck and Jeon 2019).

Among different approaches in the last decade the joint analysis of speed and ability has received increasing attention in the literature, in particular focusing on measurement aspects. Where the test is supposed to measure ability as the underlying construct for the responses, it can be assumed to measure speed as the underlying construct for the RTs as well.

The bivariate generalized linear item response theory (B-GLIRT) modelling framework, discussed by Molenaar et al. (2015), provides a broad framework for joint models. The models are essentially two-dimensional confirmatory factor analysis (CFA) models with one factor for ability and another for speed. The B-GLIRT models are measurement models, and the primary function of RTs is to enhance ability measurement. The hierarchical model (van der Linden 2007) is a prototypical model in this category and has inspired related models with different RT distributions. Roughly speaking the hierarchical model is a two-dimensional model, with one dimension for accuracy (correct vs. incorrect) interpreted as ability and another dimension for RT (log of RT) interpreted as speed. It is a measurement model with the advantage that the measurement of ability can benefit from the RT information. If the two dimensions are related, the measurement of each of them gains strength from the data for the other. The more this relationship is strong and the more the correlation structure can be captured through the proposed model, the better the process of estimating the person and the item parameters will be, even if the focus of estimation is only on one of the person parameters, such as ability (van der Linden 2007; Fox and Marianti 2016).

In order to estimate the accuracy and speed of students, we followed the approach of Fox and Marianti (2016), that is a development of the model proposed first by van der Linden (2007) and successively by Klein Entink et al. (2009a). In particular, once the data on RA based on the correct/incorrect response, and RTs are collected for each item, they are modelled following a Bayesian joint model with a hierarchical structure that, at the first level, defines separate models for responses and RTs. At

[1] Additional time is allowed to students with special needs.

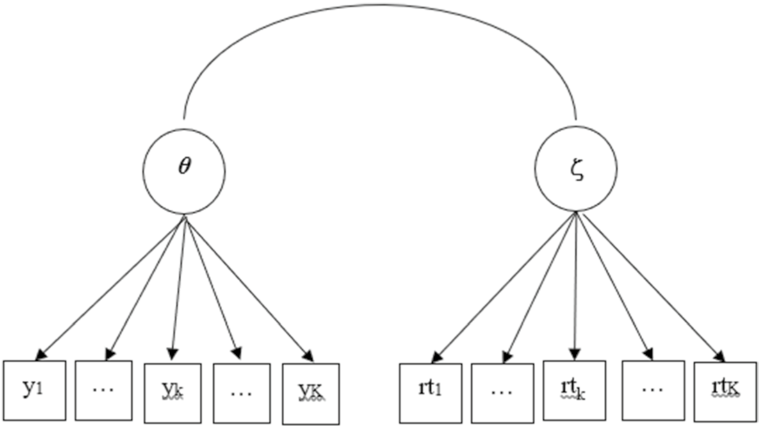

Fig. 1 Path diagram of the measurement model

the second level, a distributional structure is defined for the model parameters and hyperprior distributions are specified for the parameters.

At level 1, because only the item difficulty parameters were estimated, the one-parameter normal ogive (1PNO) model was used to define the mathematical relationship between the probability of response and the person and item parameters as follows

$$P\left(y_{ik} = 1|\theta_i, b_k\right) = \Phi\left(\theta_i - b_k\right), \tag{1}$$

where y_{ik} is the binary response variable taking value 1 when the response is correct and 0 otherwise, with $i=1,\ldots,N$ test-takers and $k=1,\ldots,K$ items, b_k is generally known as the difficulty parameter of item k, θ_i denotes the ability of test-taker i, and $\Phi(\cdot)$ is the normal cumulative distribution function.

Then, a log-normal distribution is used to model the RTs and the log RTs are stored in a $N\times K$ matrix RT. In this way, the generic element RT_{ik} is assumed to be normally distributed as follows

$$RT_{ik} = \lambda_k - \varphi_k\zeta_i + \varepsilon_{ik}, \varepsilon_{ik} \sim N\left(0, \sigma^2_{\varepsilon_k}\right), \tag{2}$$

where λ_k is the time-intensity parameter of item k, representing the population-average time (on a logarithmic scale) needed to complete the item, ζ_i is the speed parameter of test-taker i, representing the constant working speed of that test-taker, as the systematic differences in RTs given λ_k, φ_k is the time-discrimination parameter of item k, representing the sensitivity of the item for different speed levels of the test-takers. The parameter φ_k controls the decrease in expected response time on an item for a one unit increase in speed of a test-taker. Lastly,ε_{ik} is an additional error term that can model variations in RTs that cannot be explained only by the structural mean term, such as when test-takers operate with different speed values, take small pauses during the test, or change their time management. In Fig. 1, the measurement model for ability and speed is represented graphically through a path diagram. The model has a simple structure, in the sense that the binary item responses load on the ability latent variable, while the item response times load on the speed factor.

At level 2, a distributional structure is defined for the level 1 parameters. This structure is defined for both person and item parameters. For the ability and speed, a bivariate normal distribution is defined where, without identification restrictions, the hyperprior for the covariance matrix is an inverse-Wishart distribution. In the same way, a multivariate normal distribution is specified for all the item parameters of the response and response-time models, where a normal inverse-Wishart distribution is chosen as hyperprior for the mean vector and the covariance matrix.

Model parameters are estimated through the Gibbs sampling algorithm, where parameters are divided into blocks, and the simulation procedure works by iterative sampling of the conditional posterior distributions of the parameters in each block given the previous draws for the parameters in all other blocks. To identify the model, some restrictions are imposed, both for person and item parameters. For the person parameters, the mean of the ability is fixed to zero ($\mu_\theta = 0$), as well as the mean of the speed ($\mu_\zeta = 0$). As regards the item parameters, the product of the time discrimination is fixed to one $\prod\limits_k(\varphi_k) = 1$. In this way, the variance of the latent scales is fixed without restricting the variance of person parameters. In fact, by restricting the variance of a (random) person parameter (i.e. $\sigma_\theta^2 = 1$), also the covariance matrix for the person parameters will be restricted, and the inverse-Wishart distribution does not apply to a restricted covariance matrix. (Fox et al. 2021).

4.2 Bivariate multilevel regression model

Predictors of students' speed and ability were investigated through bivariate multilevel modelling (MLM), which explicitly recognizes potential correlations between the outcomes and the hierarchical data structure, namely students nested into classes. Following Rasbash et al. (2017), bivariate MLMs were specified by treating the individual student as a level 2 unit and the within-student measurements (ability and speed) as level 1 units. Moreover, in the INVALSI database, students are clustered into classes, which were specified in the MLMs as level 3 units.

There is no level 1 variation specified because level 1 exists solely to define the multivariate structure. The level 2 variances and covariance are the (residual) between-student variances. In the case where only the intercept dummy variables are fitted, and in the case where every student has both scores, the model estimates of these parameters become the usual between-student estimates of the variances and covariance. In our models, the level 3 (class-level) variances and covariances collect the contribution from unobserved contextual factors at class and higher hierarchical levels following the approach by Grilli et al. (2016). To enhance the interpretability of the results, we standardized the continuous covariates and the dependent variables (the Rasch mathematics ability and the speed estimated by the joint model introduced previously).

A sequency of bivariate MLMs was fitted to the data by iterative generalised least squares by using the software MLwiN version 3.05 (Charlton et al. 2020). First, we specified a bivariate random intercept empty model (M0), which allowed us to explore the correlations between ability and speed at class and student levels and to investigate how much variation of the response variables is present at levels 2 and 3. The empty models M0 can be written as:

$$\theta_{ij} = \beta_{0ij}\ D\_\text{ability}_{ij} \tag{3a}$$

$$\beta_{0ij} = \beta_0 + v_{0j} + u_{0ij} \tag{3b}$$

$$\zeta_{ij} = \beta_{1ij}\ D\_\text{speed}_{ij} \tag{4a}$$

$$\beta_{1ij} = \beta_1 + v_{1j} + u_{1ij}, \tag{4b}$$

where θ_{ij} and ζ_{ij} are, respectively, the ability (the standardized Rasch mathematics ability) and the speed score for student i in class j, with $i=1,\ldots, N$ test-takers, and $j=1,\ldots, J$ classes, and

$$D\_\text{ability}_{ij} = \left\{ \begin{matrix} 1 & if\ \ ability \\ 0 & if\ \ speed \end{matrix} \right\}.$$

$$D\_\text{speed}_{ij} = \left\{ \begin{matrix} 1 & if\ \ speed \\ 0 & if\ \ ability \end{matrix} \right\}.$$

The error terms at second and third levels are (u_{0ij}; u_{1ij}) and (v_{0j}; v_{1j}), respectively, with the following assumptions about means and variation

$$\begin{bmatrix} u_{0ij} \\ u_{1ij} \end{bmatrix} \sim N\left(\underline{0}, \Sigma_U\right), \Sigma_U = \begin{pmatrix} \sigma^2_{\text{u0}} & \sigma^2_{\text{u01}} \\ \sigma^2_{\text{u01}} & \sigma^2_{\text{u1}} \end{pmatrix} \tag{5}$$

$$\begin{bmatrix} \text{v}_{0\text{j}} \\ \text{v}_{1\text{j}} \end{bmatrix} \sim N\left(\underline{0}, \Sigma_V\right), \Sigma_V = \begin{pmatrix} \sigma^2_{\text{v0}} & \sigma^2_{\text{v01}} \\ \sigma^2_{\text{v01}} & \sigma^2_{\text{v1}} \end{pmatrix} \tag{6}$$

From the variance–covariance matrices it is possible to calculate the correlation between the two response variables (ability and speed) at student level and at class level. For both response variables the intra-class correlation (ICC) can be evaluated. This coefficient indicates if a multilevel approach is preferable. Successively, bivariate multilevel models including covariates of level 2 or level 3 were fitted to the data by maximum likelihood.

We included in the fixed part of model M1 the following students' covariates: gender (female, male), immigrant background (first generation, second generation, native), students' career (student repeating one or more grades, regular student), ESCS (an indicator calculated by INVALSI to evaluate the socio-economic and cultural background standardized on the gran mean and standard deviation). We also added the final mark obtained by students at the First-cycle State Leaving Examination as an index of prior achievement (final mark equal or above the national median; final mark below the national median). From Eqs. (3a), (3b) and (4a), (4b) we wrote the model M1 as follows

$$\theta_{ij} = \beta_0 + \boldsymbol{\beta}_{ab} X_{ij} + v_{0j} + u_{0ij} \tag{7}$$

$$\zeta_{ij} = \beta_1 + \boldsymbol{\beta}_{sp} X_{ij} + v_{1j} + u_{1ij}, \tag{8}$$

where $\mathbf{X}$ are the students' covariates while $\boldsymbol{\beta}_{ab}$ and $\boldsymbol{\beta}_{sp}$ are the regression coefficients for ability and speed, respectively. In this model, only the intercepts are random at the class level. The random effects assume the same formulation of M0.

Then we estimated the model M2 including also the polynomial fixed effects (linear + quadratic + cubic) of the student anxiety score. This score is based on the student responses to four ordinal items formulated to evaluate the emotional status of test-takers and administered at the end of the test. Following an item response theory (IRT) approach, the graded response model (Samejima 1969) was used to estimate a continuous student anxiety score based on the student responses on the four items (Matteucci et al. 2023).

In Model M3 covariates at classroom level (and above) were introduced.

$$\theta_{ij} = \beta_0 + \boldsymbol{\beta}_{ab} X_{ij} + \boldsymbol{\gamma}_{ab} W_j + v_{0j} + u_{0ij} \tag{9}$$

$$\zeta_{ij} = \beta_1 + \boldsymbol{\beta}_{sp} X_{ij} + \boldsymbol{\gamma}_{sp} W_j + v_{1j} + u_{1ij} \tag{10}$$

The following class-compositional variables were included: percentage of students with an immigrant background, students repeating one or more grades, students with a low final mark at the end of the First-cycle State Leaving Examination. We also added the class average ESCS and the average anxiety score. The school track (lyceum, technical institute, vocational) and the geographical area (North-East, North-West, Center, South and the Islands) were also introduced in M3.

5 Results and discussion

5.1 First step: estimating item parameters, ability and speed

As regards the joint modelling of RA and RTs, the analyses were done by using the R package LNIRT (Fox et al. 2021). The average population level of speed was fixed to zero to identify the scale. The main results for the item parameters are summarized in Table 1.

The estimated mean of the item difficulties was − 0.070 and the range was relatively large, which gives support to accurate estimation of test-takers' ability. The statistics on the item difficulty were slightly different from the ones reported in the official INVALSI report (see Desimoni 2019). In fact, the item difficulty parameters were estimated by INVALSI during the item calibration phase, while we re-estimated them with the available data after test administration. Note that the LNIRT package uses the 1PNO model (1), while the model assumed for calibration was the

Table 1 Item parameters

	Item difficulty (Rasch model)	Time intensity	Time discrimination
Mean	−0.070	4.229	1.175
Minimum	−2.574	3.114	0.011
Maximum	2.726	5.151	2.288

Table 2 Person parameters

	Person ability	Person speed
Mean	0.000	0.000
Minimum	−2.311	0.611
Maximum	1.946	2.283

Table 3 Correlation matrix for item parameters (*p*-values in brackets)

	Item difficulty	Time intensity	Time discrimination
Item difficulty	1.000	0.370 (0.000)	0.234 (0.004)
Time intensity	0.370 (0.000)	1.000	−0.014 (0.436)
Time discrimination	0.234 (0.004)	−0.014 (0.436)	1.000

Table 4 Correlation matrix for person parameters (*p*-values in brackets)

	Person ability	Person speed
Person ability	1.000	−0.574 (0.000)
Person speed	−0.574 (0.000)	1.000

Rasch model. For this reason, to compare the two estimates, it was first necessary to multiply by 1.7 those provided by the package (Fox et al. 2021). The estimated mean of the time intensities was around 4.2 and the time intensities ranged from 3.11 to 5.15. So, the average RT to complete each item ranged from exp(3.11)≈22 to exp(5.15)≈172 s. The estimated mean time discrimination was 1.17 with a quite high variability (0.01 to 2.29 on a logarithmic scale), indicating that the items discriminated substantially between test-takers of different speed.

For person parameters, the estimates of ability and speed are given in Table 2. The ability followed a normal distribution, while the speed distribution curve was slightly skewed. From the residual analysis, it turned out that the residuals of the response times violated the assumption of log-normal distribution for most items. Following several analyses, it was possible to note that this violation was due to the large number of test-takers (n = 35,970).

The correlation matrices for item parameters and for person parameters are illustrated in Tables 3 and 4, respectively.

The analysis of these results allows us to say that there was, on average, a positive relationship between the difficulty of the items and their intensity and discriminating power, in terms of time. This means that the most difficult (easy) items were also the ones that discriminated better (worse) and required more (less) time to perform. The fact that the more difficult items tended to be more time consuming is in line with the common assumption that the more complex cognitive reasoning items require more processing steps by the test-taker.

Table 4 provides important information about the correlation between the speed and ability of the test-takers (−0.574), which was negative and significant. So, test-takers with a higher (lower) ability tended to be slower (faster). This result is consistent with previous findings on the between-person speed-ability association in the mathematics domain (Costa and Chen 2023; Zhang et al. 2024), as well as in other task domains that require controlled cognitive processing (Goldhammer et al. 2014; Naumann and Goldhammer 2017), when test-takers are not under the pressure of strict time limits (Klein Entink et al. 2009b).

Finally, the extreme residual analysis gave the following results: around 15.54% of RT patterns were considered extreme with 95% posterior probability, while for the RA patterns the percentage was 2.19%. When considering the joint pattern (RA and RT), only 0.49% of these were extremes. The residual variance was around 0.488 and the variance in working speed and time intensities were not so small. Therefore, RT outliers only slightly affected the fit of the log-normal distribution, going to confirm what has already been anticipated that the violation of the log-normality was caused only by the large number of test-takers.

5.2 Second step: investigating the covariates of ability and speed by considering the hierarchical data structure

As for the MLM results, the empty model M0 showed that the correlation between the mathematics ability and the speed at the student level was negative (−0.484): the high-ability test-takers worked slower on computer-based items than the low-ability test-takers. The correlation at the class level was also negative (−0.779) but higher than the correlation at the student level. The estimated intraclass correlation coefficients (ICCs) indicated that ability scores of students in the same classroom were correlated (class ICC = 0.53); a similar result emerged for speed scores (class ICC = 0.48). Therefore, a multilevel bivariate approach seems to be appropriate for representing the structure of the data.

The other three models were then estimated. Table 5 summarizes the results from the likelihood ratio test (LRT) to choose the preferable one.

Results from model comparison suggest M3 as the final model showing both the smallest indicator (−2*loglikelihood) and a significant LRT. In Tables 6 and 7 the results from M3 are reported.

We distinguished the covariates with respect to level 2 (student) and level 3 (class and above). Controlling for all other predictors, we found that students with low prior achievement were less accurate and spend less time on mathematics items than their peers. A similar pattern of results emerged for the fixed effect

Table 5 Model comparison via LRT

Model	– 2*Loglikelihood	Comparison	LR χ^2	d.f	*p*-value
M0	156,145.167				
M1	150,742.696	M1–M0	5402.471	12	< 0.0001
M2	147,629.961	M2–M1	3112.735	6	< 0.0001
M3	145,392.581	M3–M2	2237.38	22	< 0.0001

Table 6 Final model parameter estimates: fixed effects

	Ability			Speed		
	Estimate	S.E	*p*-value	Estimate	S.E	*p*-value
Fixed effects (*β*)						
Intercept	0.507	0.05	0.000	– 0.377	0.069	0.000
Male vs. female	0.111	0.008	0.000	0.091	0.009	0.000
Student's ESCS	0.001	0.004	0.732	0.019	0.005	0.000
Student repeating one or more grades vs. not	– 0.149	0.011	0.000	0.217	0.012	0.000
Low prior achievement vs. average and high	– 0.442	0.008	0.000	0.253	0.01	0.000
Second-generation immigrant vs. native	– 0.084	0.016	0.000	0.006	0.018	0.747
First-generation immigrant vs native	– 0.089	0.016	0.000	– 0.058	0.019	0.002
Math test anxiety	– 0.176	0.007	0.000	– 0.07	0.008	0.000
Math test anxiety^2	– 0.007	0.004	0.046	0.073	0.004	0.000
Math test anxiety^3	0.007	0.003	0.010	0.011	0.003	0.000
Class % of stud. with low prior achievement	– 0.006	0.001	0.000	0.004	0.001	0.000
Class % of immigrants	– 0.005	0.001	0.000	0.006	0.001	0.000
Class average ESCS	0.226	0.029	0.000	– 0.175	0.04	0.000
Class % of students repeating grades	– 0.001	0.001	0.109	0.003	0.001	0.003
Class average math test anxiety	– 0.043	0.026	0.089	– 0.29	0.035	0.000
North West vs. Center	0.212	0.028	0.000	– 0.171	0.038	0.000
North East vs. Center	0.251	0.028	0.000	– 0.235	0.038	0.000
South vs center	– 0.217	0.029	0.000	0.118	0.04	0.004
South Islands vs. Center	– 0.436	0.029	0.000	0.296	0.04	0.000
Liceum vs vocational	0.109	0.038	0.004	– 0.245	0.052	0.000
Technical vs. Vocational	0.177	0.027	0.000	– 0.362	0.037	0.000

of being a student who repeated one or more grades. The higher speed for students with a history of struggling academically may be indicative of avoidance motivation, lack of effort, and the desire to withdraw (Sideridis and Alahmadi 2022). As mentioned in the literature review, past experiences of failure or success are likely to shape students expectations of success (Eccles and Wigfield 2002, 2020), which in turn may influence their willingness to strive and persist in the testing situation, especially under low-stakes conditions.

Table 7 Final model parameter estimates: random effects

Random effects
Between-class cov. Matrix Σ_v
$\begin{array}{rcc} & \text{Ability} & \text{Speed} \\ \Sigma_v = & \left(\begin{matrix} 0.142 \\ -0.147 \end{matrix}\right. & \left.\begin{matrix} -0.147 \\ 0.285 \end{matrix}\right) \end{array}$
Correlation −0.727
Within-class cov. Matrix Σ_u
$\begin{array}{rcc} & \text{Ability} & \text{Speed} \\ \Sigma_u = & \left(\begin{matrix} 0.401 \\ -0.225 \end{matrix}\right. & \left.\begin{matrix} -0.225 \\ 0.522 \end{matrix}\right) \end{array}$
Correlation −0.492

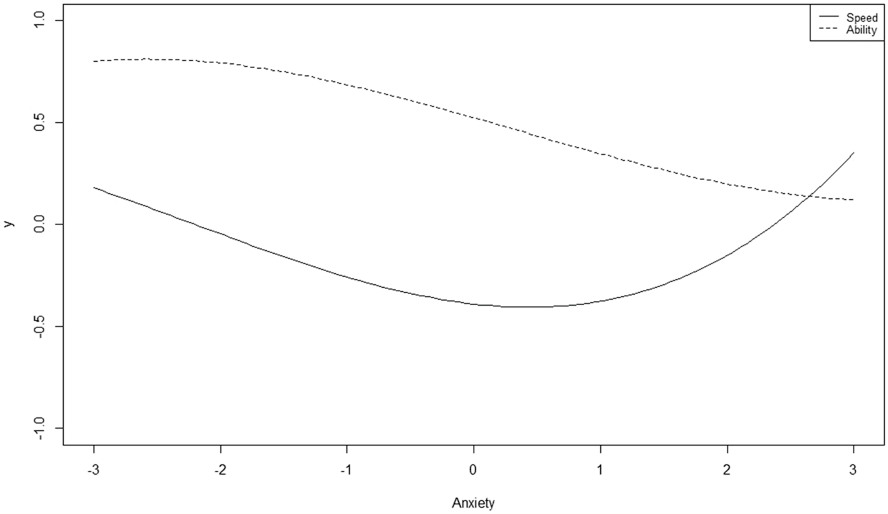

Fig. 2 Relation of ability and speed to anxiety

Students' self-reported anxiety before and during the test was related to ability and speed in a non-linear fashion (Fig. 2).

Ability estimate reached its local maximum (≈ 0.794) at very low level of test anxiety (x ≈ −2.58) and changed little through the lowest levels of test anxiety. Moving from low-intermediate to high-intermediate level of test anxiety, ability decreased substantially and then levelled off, reaching its local minimum (0.101) at very high level of test anxiety (x ≈ 3.247). A different pattern of results emerged for speed. For students experiencing low to average levels of test anxiety, speed consistently decreased as the anxiety increased, and then levelled off (local maximum ≈ 0.425 at x ≈ −4.86; local minimum ≈ −0.392 at x ≈ 0.463). For the higher levels of text anxiety, the speed increased sharply as the anxiety increased. These results are consistent with previous findings suggesting a nonlinear relationship between test anxiety and other variables (Cassady and Finch 2020; Jerrim 2022).

The decrease in test-taking speed from the lower to the intermediate levels of test anxiety is consistent with the PET (Eysenck and Calvo 1992) and the ACT (Eysenck et al. 2007) predictions that increased anxiety leads to the allocation of additional effort to attempt to maintain task performance. However, for the higher levels of test anxiety, a different pattern emerged that is plausible due to students' willingness to finish the test quickly at the expense of the response accuracy and their effective ability estimate. The latter results are consistent with those observed by Akhtar and Firdiyanti (2023).

As for gender, the unique associations with speed and ability were both positive and very similar in size: males were slightly more accurate and worked slightly faster than females. This result is in contrast from Costa and Chen (2023), who found no gender difference in test-taking speed on a subsample of mathematics items from PISA 2012. However, it is consistent with other previous findings on RTs, such as those observed by Kapoor et al. (2023) on PISA 2018 data, and with previous research on gender differences in test-taking behaviour under low-stakes conditions (Balart and Oosterveen 2019; Azzolini et al. 2019; Teig et al. 2020; Rutkowski et al. 2023; Marcq and Braeken 2023). Despite slower items processing, female students performed worse than their male counterparts. This could suggest that the processing of mathematics items is more efficient for male students than for female students.

The socio-economic and cultural background indicator deserves a special attention. The unique effect of students' ESCS on ability was not statistically significant, whilst a weak, albeit significant, positive effect emerged for speed. It is important to note that M3 also included the direct effect of prior achievement. Therefore, we cannot exclude the possibility that the ESCS influences mathematics ability indirectly, through its effect on prior achievement (or other covariates). After accounting for individual differences in past achievement-related experiences, there were very slight differences based on students' family background, with those from more affluent families completing the test slightly more efficiently. Native students outperformed immigrant students, and first-generation immigrants performed slightly, but significantly, slower than natives. This finding contrasts with that of Costa and Chen (2023), who found no differences in test-taking speed between Scandinavian students with different immigrant backgrounds. A possible explanation for our result is that first-generation immigrants, especially those who are non-native speakers of Italian, are less fluent in reading the items and thus may need more time to process the items and have a higher propensity to revisit the items. This is consistent with the findings of Bezirhan et al. (2021) on test-taking behaviour in a different testing environment.

After controlling for relevant individual-level predictors, school type and the geographical area of the school, the contextual effect of class ESCS on ability and speed was significant: students from classes with higher ESCS spent more time on items and obtained better results in terms of ability. The percentage of students with an immigrant background was associated with lower ability and higher speed; analogous results emerged for the percentage of students with low prior achievement. Students attending classes with higher average test-related anxiety spent more time on items. Our results corroborate earlier findings showing that individual students' performance in achievement tests are related to the composition of their learning environment (see meta-analysis by van Ewijk and Sleegers 2010) and provide novel evidence on class-level predictors of individual test taking speed. The observed effects of classroom composition on speed may be partly explained by peer effects on test-taking motivation and/or socio-emotional status. For example, in line with the EVT (Eccles and Wigfield 2002, 2020), peers with high socio-economic status may share beliefs and values that are positively associated with engagement and effort to achieve the best possible result in an achievement test. Therefore, individual students in classes with high average ESCS may be more likely to increase the effort invested in the test by decreasing the rate at which they work. In addition, how the peers group feels about testing situation may influence individual students' test taking behaviour: for example, students whose peers are very anxious and worried about taking the test may be more cautious when responding and take precautionary measures (e.g., double-checking) that would decrease their speed.

Significant differences in ability and speed also emerged by school tracks and geographical area. Students from the vocational school were less accurate and spent less time on the items than those from the lyceum and technical institute. Students from the North-East and the North-West were more accurate and worked slowly on items than those from the Center of Italy, whilst those from the South and the South and Islands were less accurate and spent less time on items. The random effects confirmed a negative correlation at both levels but higher between class.

6 Concluding remarks

In recent years, there has been an increased interest for the between-person speed-ability association in low-stakes LSAs. The present study contributed to this stream of research by applying one of the most popular models for responses and response times in tests, i.e. van der Linden's (2007) hierarchical model, to the process and product data from the INVALSI national assessment of students' achievement in mathematics. The results showed a negative association across test-takers between their effective working speed on items and their effective mathematics ability. The direction of this between-person association is consistent with previous findings on mathematics assessments (Costa and Chen 2023; Zhang et al. 2024) and supports the idea that when the task domain requires deliberate, controlled cognitive processing, less time on task is associated with better performance (Goldhammer et al. 2014; Naumann and Goldhammer 2017).

Moreover, utilizing a bivariate multilevel regression model, which incorporates individual explanatory variables, allowed for a comprehensive examination of how individual characteristics influence test results, accounting for both speed and ability estimates. This approach proves effective in addressing the complexities of hierarchical educational data, revealing significant differences in student performance based on various factors such as prior achievement, math test anxiety, sociodemographic characteristics, class composition, school tracks, and geographical area.

Other approaches (Fox and Marianti 2016; Bolsinova et al. 2017) have focused on including regression structures to examine the effects of covariates on ability and speed differences among groups but these solutions do not take into account the multilevel structure of the data. Another family of IRT models, the explanatory item response models (de Boeck and Wilson 2004), decomposes common variability across item and person clusters using person properties, potentially reducing uncertainty but failing to address the challenge of combining speed and ability.

However, the paper acknowledges several limitations that warrant consideration for future research. Firstly, the assumption of log-normality of the RT distribution should be verified, exploring alternative distributions and identifying aberrant response time patterns that could affect estimation accuracy. Additionally, the implicit assumption that the speed parameter remains constant throughout the test may not hold true, as test-takers may experience changes in speed due to factors such as fatigue or adopting new strategies. Exploring models where response time is influenced by both speed and ability, such as within-item multidimensionality models, could provide valuable insights.

In conclusion, this study lays a significant foundation for future research utilizing data from online administration tests to enhance and refine the assessment of learning outcomes.

Acknowledgements The authors would like to thank INVALSI for granting access to the data. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of INVALSI.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. This study was partially funded by the European Union–NextGenerationEU, in the framework of the "GRINS–Growing Resilient, INclusive and Sustainable project" (PNRR–M4C2–I1.3–PE00000018–CUP J33C22002910001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

Akhtar H, Firdiyanti R (2023) Test-taking motivation and performance: do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? Learn Individ Differ 106:102323. https://doi.org/10.1016/j.lindif.2023.102323

Azzolini D, Bazoli N, Lievore I, Schizzerotto A, Vergolini L (2019) Beyond achievement. A comparative look into 15-year-olds' school engagement, effort and perseverance in the European Union. Publication Office of the European Union

Balart P, Oosterveen M (2019) Females show more sustained performance during test-taking than males. Nat Commun 10(1):3798. https://doi.org/10.1038/s41467-019-11691-y

Bandura A (1997) Self-efficacy: the exercise of control. Freeman, New York

Bezirhan U, von Davier M, Grabovsky I (2021) Modeling item revisit behavior: the hierarchical speed–accuracy–revisits model. Educ Psychol Measur 81(2):363–387. https://doi.org/10.1177/0013164420950556

Bolsinova M, de Boeck P, Tijmstra J (2017) Modelling conditional dependence between response time and accuracy. Psychometrika 82:1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bolsinova M, Tijmstra J (2018) Improving precision of ability estimation: getting more from response times. Br J Math Stat Psychol 71(1):13–38. https://doi.org/10.1111/bmsp.12104

Cassady JC, Finch WH (2020) Revealing nuanced relationships among cognitive test anxiety, motivation, and self-regulation through curvilinear analyses. Front Psychol 11:1141. https://doi.org/10.3389/fpsyg.2020.01141

Charlton C, Rasbash J, Browne WJ, Healy M, Cameron B (2020) MLwiN version 3.05. Centre for multilevel modelling. University of Bristol, Bristol

Costa DR, Chen C-W (2023) Exploring the relationship between process data and contextual variables among Scandinavian students on PISA 2012 mathematics tasks. Large-Scale Assess Educ 11(1):5. https://doi.org/10.1186/s40536-023-00155-x

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

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

De Boeck P, Scalise K (2019) Collaborative problem solving: processing actions, time, and performance. Front Psychol 10:1280. https://doi.org/10.3389/fpsyg.2019.01280

DeMars CE, Bashkov BM, Socha AB (2013) The role of gender in test-taking motivation under low-stakes conditions. Res Pract Assess 8:69–82

Desimoni M (2019) Le prove computer based per la rilevazione nazionale degli apprendimenti INVALSI 2018: aspetti metodologici. Retrieved from https://invalsi-areaprove.cineca.it/docs/2019/Parte_I_capitolo_2_aspetti_metodologici_CBT_2018.pdf

Eccles JS, Adler TF, Futterman R, Goff SB, Kaczala CM, Meece JL, Midgley C (1983) Expectancies, values, and academic behaviors. In: Spence JT (ed) Achievement and achievement motivation. W. H. Freeman, San Francisco, pp 75–146

Eccles JS, Wigfield A (2002) Motivational beliefs, values, and goals. Annu Rev Psychol 53(1):109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eccles JS, Wigfield A (2020) From expectancy-value theory to situated expectancy-value theory: a developmental, social cognitive, and sociocultural perspective on motivation. Contemp Educ Psychol 61:101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Eysenck MW, Calvo MG (1992) Anxiety and performance: the processing efficiency theory. Cogn Emot 6(6):409–434. https://doi.org/10.1080/02699939208409696

Eysenck MW, Derakshan N, Santos R, Calvo MG (2007) Anxiety and cognitive performance: attentional control theory. Emotion 7(2):336–353. https://doi.org/10.1037/1528-3542.7.2.336

Fox J-P, Klein Entink R, van der Linden WJ (2007) Modeling of responses and response times with the package cirt. J Stat Softw 20(7):1–14. https://doi.org/10.18637/jss.v020.i07

Fox J-P, Marianti S (2016) Joint modeling of ability and differential speed using responses and response times. Multivar Behav Res 51(4):540–553. https://doi.org/10.1080/00273171.2016.1171128

Fox J-P, Klotzke K, Simsek AS (2021) LNIRT: an R package for joint modeling of response accuracy and times. arXiv preprint. https://doi.org/10.48550/arXiv.2106.10144

Goldhammer F, Naumann J, Stelter A, Tóth K, Rölke H, Klieme E (2014) The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. J Educ Psychol 106(3):608–626. https://doi.org/10.1037/a0034716

Grilli L, Pennoni F, Rampichini C, Romeo I (2016) Exploiting TIMSS and PIRLS combined data: multivariate multilevel modelling of student achievement. Ann Appl Stat 10(4):2405–2426. https://doi.org/10.1214/16-AOAS988

Hansen Y, Radišić J, Ding Y, Liu X (2022) Contextual effects on students' achievement and academic self-concept in the Nordic and Chinese educational systems. Large-Scale Assess Educ 10(1):16. https://doi.org/10.1186/s40536-022-00133-9

INVALSI (2018a) Quadro di riferimento delle prove invalsi di matematica. Retrieved from https://invalsi-areaprove.cineca.it/docs/file/QdR_MATEMATICA.pdf

INVALSI (2018b) Rapporto prove INVALSI 2018. Retrieved from https://www.invalsi.it/invalsi/doc_evidenza/2018/Rapporto_prove_INVALSI_2018.pdf

Jerrim J (2022) The power of positive emotions? The link between young people's positive and negative affect and performance in high-stakes examinations. Assess Educ Princ, Policy Pract 29(3):310–331. https://doi.org/10.1080/0969594X.2022.2054

Kapoor R, Fahle E, Kanopka K, Klinowski D, Ribeiro ACT, Domingue B (2023) Differences in time usage as a competing hypothesis for observed group differences in accuracy with an application to observed gender differences in PISA data. PsyArXiv. https://psyarxiv.com/6wsmq/download/?format=pdf

Ketonen EE, Hotulainen R (2019) Development of low-stakes mathematics and literacy test scores during lower secondary school–a multilevel pattern-centered analysis of student and classroom differences. Contemp Educ Psychol 59:101793. https://doi.org/10.1016/j.cedpsych.2019.101793

Klein Entink RH, Fox J-P, van der Linden WJ (2009a) A multivariate multilevel approach to the modeling of accuracy and speed of test takers. Psychometrika 74(1):21–48. https://doi.org/10.1007/s11336-008-9075-y

Klein Entink RH, Kuhn J-T, Hornke LF, Fox J-P (2009b) Evaluating cognitive theory: a joint modeling approach using responses and response times. Psychol Methods 14(1):54–75. https://doi.org/10.1037/a0014877

Kroehne U, Hahnel C, Goldhammer F (2019) Invariance of the response processes between gender and modes in an assessment of reading. Front Appl Math Stat 5:2. https://doi.org/10.3389/fams.2019.00002

Marcq K, Braeken J (2023) Gender differences in item nonresponse in the PISA 2018 student questionnaire. Educ Assess Eval Account. https://doi.org/10.1007/s11092-023-09412-7

Matteucci M, Mignani S, Spaccapanico Proietti G (2023) Some insights on the relationship between student performance and test related emotional aspects. In: Falzetti P (ed) The school and its protagonists: the students. V seminar "INVALSI data: a tool for teaching and scientific research. Franco Angeli, Milano, pp 26–48

Molenaar D, Tuerlinckx F, van der Maas HL (2015) A bivariate generalized linear item response theory modelling framework to the analysis of responses and response times. Multivar Behav Res 50(1):56–74. https://doi.org/10.1080/00273171.2014.962684

Naglieri JA, Rojahn J (2001) Gender differences in planning, attention, simultaneous, and successive (PASS) cognitive processes and achievement. J Educ Psychol 93(2):430

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

Pressley M, Borkwski JG, Schneider W (1989) Good information processing: what it is and how education can promote it. Int J Educ Res 13(8):857–867. https://doi.org/10.1016/0883-0355(89)90069-4

Rasbash J, Steele F, Browne WJ, Goldstein H (2017) A user's guide to MLwiN, v3.00. Centre for Multilevel Modelling. University of Bristol, Bristol

Reis Costa D, Leoncio Netto W (2022) Process data analysis in ILSAs. In: Nilsen T, Stancel-Piątak A, Gustafsson J-E (eds) International handbook of comparative large-scale studies in education: perspectives, methods and findings. Springer International Publishing, Berlin, pp 1–27. https://doi.org/10.1007/978-3-030-38298-8_60-1

Rutkowski D, Rutkowski L, Valdivia DS, Canbolat Y, Underhill S (2023) A census-level, multi-grade analysis of the association between testing time, breaks, and achievement. Appl Measur Educ 36(1):14–30. https://doi.org/10.1080/08957347.2023.2172019

Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. Psychometrika 34:1–97. https://doi.org/10.1007/BF03372160

Schneider W, Shiffrin RM (1977) Controlled and automatic human information processing: I. Detection, search, and attention. Psychol Rev 84(1):1–66

Shiffrin RM, Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychol Rev 84(2):127–190

Sideridis G, Alahmadi MTS (2022) The role of response times on the measurement of mental ability. Front Psychol 13:892317. https://doi.org/10.3389/fpsyg.2022.892317

Skaalvik EM (1997) Self-enhancing and self-defeating ego orientation: relations with task and avoidance orientation, achievement, self-perceptions, and anxiety. J Educ Psychol 89(1):71–81. https://doi.org/10.1037/0022-0663.89.1.71

Teig N, Scherer R, Kjærnsli M (2020) Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data. J Res Sci Teach 57(9):1400–1429. https://doi.org/10.1002/tea.21657 van der Linden WJ (2007) A hierarchical framework for modeling speed and accuracy on test items. Psychometrika 72(3):287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden WJ (2009) Conceptual issues in response-time modeling. J Educ Meas 46(3):247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

van der Linden WJ, Klein Entink RH, Fox J-P (2010) IRT parameter estimation with response times as collateral information. Appl Psychol Meas 34(5):327–347. https://doi.org/10.1177/014662160934

van der Linden WJ, Fox J-P (2016) Joint hierarchical modeling of responses and response times. In: van der Linden WJ (ed) Handbook of item response theory, volume one models. Chapman and Hall/CRC, New York, pp 481–500

van Ewijk R, Sleegers P (2010) The effect of peer socioeconomic status on student achievement: a meta-analysis. Educ Res Rev 5(2):134–150. https://doi.org/10.1016/j.edurev.2010.02.001

von der Embse N, Jester D, Roy D, Post J (2018) Test anxiety effects, predictors, and correlates: a 30-year meta-analytic review. J Affect Disord 227:483–493. https://doi.org/10.1016/j.jad.2017.11.048

Wigfield A, Eccles JS (2000) Expectancy–value theory of achievement motivation. Contemp Educ Psychol 25(1):68–81. https://doi.org/10.1006/ceps.1999.1015

Winne PH, Hadwin AF (1998) Studying as self-regulated learning. Metacognition in educational theory and practice. Lawrence Erlbaum Associates Publishers, Mahwah, pp 277–304

Zhang M, Andersson B, Jin S (2024) Fast estimation of generalized linear latent variable models for performance and process data with ordinal, continuous, and count observed variables. Br J Math Stat Psychol. https://doi.org/10.1111/bmsp.12337

Zimmerman ME (2011) Speed–accuracy tradeoff. In: Kreutzer JS, DeLuca J, Caplan B (eds) Encyclopedia of clinical neuropsychology. Springer, New York, pp 2344–2344. https://doi.org/10.1007/978-0-387-79948-3_1247

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