The joint estimation of accuracy and speed: An application to the INVALSI data

Luca Bungaro, Marta Desimoni, Mariagiulia Matteucci, Stefania Mignani

1. Introduction

In recent years, the implementation of computer based testing (CBT) has been receiving a growing interest because of its operational advantages. CBT allows to automatically collect data not only on the students' response accuracy (RA) based on item responses, but also on their response times (RT). Using the RTs, the assessment results can be further improved in terms of precision, fairness, and minimizing costs. The information obtained by RTs can be used for item calibration, test design, detection of cheating, and adaptive item selection.

The RTs used to respond to items provide information about working speed, where RA data provide information about ability. RTs are collected for estimating speed and item time-intensity (i.e., population-average amount of time needed to complete an item), to investigate relationships with speed components and accuracy, but also to investigate several issues in educational testing.

In Italy, the National Institute for the Evaluation of the Education and Training System (INVALSI) every year administers standardized tests via CBT to students attending grades 8, 10, and 13. In this study, we use the 2018 mathematics data for grade 10 to estimate the ability and speed of students and to evaluate the impact of some students' characteristics both to the performance and to the response time behaviour.

In the INVALSI test the number of involved examinees is very large and tests must be administered in multiple sessions and locations. Moreover, testing organizations 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 a Rasch item bank through automated test assembly.

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 minutes on grade 10 tests, which is considered enough for students to read and answer all the questions[1]. These time constraints may have had an impact on the speed that must be considered in the results' discussion.

In the first step of the analysis, we implemented the fully Bayesian approach of Fox *et al.* (2021), following the models of van der Linden (2007) and Klein Entik *et al.* (2009). In the second step, considering the hierarchical nature of the data, we use the estimated mathematics ability and speed in a bivariate multilevel model, where the first-level units are represented by students and the second-level units are represented by classes. Covariates such as gender, school type, immigrant status, economic, social, and cultural status, prior achievement, grade retention, student anxiety, class compositional variables, and geographical area are included in the model.

2. Methods

The models for estimating the accuracy and speed of students and for investigating the relation

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

Luca Bungaro, University of Bologna, Italy, luca.bungaro2@unibo.it
Marta Desimoni, INVALSI, Italy, marta.desimoni@invalsi.it, 0000-0002-3407-0002
Mariagiulia Matteucci, University of Bologna, Italy, m.matteucci@unibo.it, 0000-0003-3404-6325
Stefania Mignani, University of Bologna, Italy, stefania.mignani@unibo.it, 0000-0003-4746-1130

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Luca Bungaro, Marta Desimoni, Mariagiulia Matteucci, Stefania Mignani, *The joint estimation of accuracy and speed: An application to the INVALSI data*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.39, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 221-226, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

between these outcome variables and a set of predictors are described in the following.

2.1 Models for responses and response times

In order to estimate the accuracy and speed of students, we followed the approach of Fox *et al*. (2021), who implemented in the `R` package `LNIRT` the models of van der Linden (2007) and Klein Entik *et al.* (2009). In particular, once the data on RA, i.e. 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 response times. At the second level, a distributional structure is defined for the model parameters and hyperprior distributions are specified for the parameters.

At level 1, 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

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

where y_{ik} is the binary response variable taking value 1 when the response is correct and 0 otherwise, with $i = 1, ..., N$ test-takers and $k = 1, ..., K$ 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 × K matrix *RT*. In this way, the generic element RT_{ik} is assumed to be normally distributed as follows

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

where λ_k is the time-intensity parameter of item k, representing the population-average time (on a logarithmic scale) needed to complete an 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. 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.

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. As regards the item parameters, the product of the time discrimination is fixed to one $\prod_k(\varphi_k) = 1$. For the person parameters, the mean of the ability is fixed to zero, as well as the mean of the speed. In this way, the `LNIRT` package is able to avoid restricting the variance of a person parameter, which would otherwise have resulted in the restriction of the covariance matrix (for the details on model estimation and identification, see Fox *et al*., 2021).

2.2 Bivariate multilevel 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. Following Rasbash *et al.* (2017), bivariate MMs were specified by treating the individual student as a level 2 unit (n = 35,727) and the within-student measurements (Ability and Speed) as level 1 units. Students (n = 243) with missing values in the covariates have been excluded from the MLMs data. In the INVALSI database, students are clustered into classes, which were specified in the MLMs as level 3 units (n = 2,273). In turn, classes are nested into schools. However, since in the INVALSI national sample a maximum of two classes are sampled within each school, we preferred to not fit a four-level model also including the school level. Therefore, in our models, the class-level random effects collected the unobserved contextual factors at class and higher hierarchical levels.

To enhance the interpretability of the results, we standardized the continuous covariates and the dependent variables (Rasch ability estimate and person speed estimate from `LNIRT`). The following bivariate MLMs were fitted to the data by Iterative Generalised Least Squares using `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 response variables variation is present at levels 2 and 3. Level 1 existed solely to define the bivariate structure and there was no level 1 variation specified in the bivariate MLMs (Rasbash *et al.*, 2017).

In model M1, we added to M0 the fixed effects of students' sociodemographic characteristics, prior achievement (0 = the final mark at the First-cycle State Leaving Examination is equal or above the national median; 1 = the final mark is below the national median), school career (1 = student repeating one or more grades, 0 = otherwise), and mathematics test anxiety.

In model M2, the following L2 variables were included: class average ESCS and math test anxiety; the 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.

In the final model (M3), we added the school track (two dichotomous variables: vocational vs lyceum; vocational vs technical institute, reference category = vocational) and the geographical area (4 dichotomous variables, Center vs North-West; Center vs North-East; Center vs South; Center vs South and the Islands; reference category: Center).

The likelihood-ratio (LR) test was used to compare the nested models described above (M1 vs M0; M2 vs M1; M3 vs M2).

3. Results

As regards the joint modelling of RA and RTs, the main results for item parameters are summarized in Table 1, which shows mean, minimum, and maximum of the expected a posteriori (EAP) estimates.

Table 1. Item parameters

	Item Difficulty (Rasch Model)	Time Intensity	Time Discrimination	Difficulty Difference (Rasch Model)
Mean	-0.070	4.229	1.175	0.108
Minimum	-2.574	3.114	0.011	0.001
Maximum	2.726	5.151	2.288	0.281

The last column of Table 1 shows the absolute value of the difference between the parameter b, estimated by the model, and the one obtained during the calibration of the items. Note that the `LNIRT` package uses the 1PNO model (1), while the model assumed for calibration was the Rasch model, also known as the one-parameter logistic (1PL) 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).

For person parameters, the estimates of ability and speed are given in Table 2.

Table 2. Person parameters

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

The ability follows a normal distribution, while the speed distribution curve is slightly skewed. From the residual analysis, it turns out that the residuals of the response times violate the assumption of log-normal distribution for most items. Following several analyses, it was possible to note that this violation is due to the large number of test-takers (35,970) and the very nature of the INVALSI test.

The correlation matrices for person and item parameters are given in Table 3 and Table 4, respectively. The analysis of these results allows us to say that there is, 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 are also the ones that discriminate better (worse) and require more (less) time to perform. The negative correlation between time-discrimination and time-intensity, on the other hand, indicates that on average the items that require more (less) time are the ones that discriminate worse (better), but with a very low and not significant magnitude.

Table 3. Item correlation matrix

	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 provides important information about the correlation between the speed and ability of the test-takers (-0.574), which is negative and significant. So, test-takers with a higher (lower) ability tends to be slower (faster).

Table 4. Person correlation matrix

	Person Ability	Person Speed
Person Ability	1.000	-0.574 (0.000)
Person Speed	-0.574 (0.000)	1.000

This result is known in the literature. In particular, it goes to consolidate that hypothesis for which those who are prepared want to engage and show their skills, even during a test that does not directly affect their school average, while those who are less prepared tend to be less interested and more hasty.

Finally, the extreme residual analysis gave the following results: around 15.54% of RT patterns are considered extreme with 95% posterior probability, while for the RA patterns the percentage is 2.19%. When considering the joint pattern (RA and RT), only 0.49% of these are extremes. The residual variance is around 0.488 and the variance in working speed and time intensities are not so small. Therefore, RT outliers only slightly affect the fit of the log-normal distribution, going to confirm what has already been anticipated about the nature of the test itself.

As for the MLM results, M0 shows that the high-ability test-takers worked slower on computer-based items than the low-ability test-takers (within-classes correlation = -.484). The between-classes correlation between speed and ability is higher than the correlation at the student level (-.779). The estimated intraclass correlation coefficients (ICCs) indicate that ability scores of students in the same classroom are correlated (ability: school ICC = .53); a similar result emerges for speed scores (speed: school ICC = .48). Therefore, a multilevel bivariate approach seems to be

appropriate for representing the structure of the data.

Table 5. Likelihood ratio test

Model	-2*Loglikelihood	Comparison	LR χ^2	d.f.	*p*-value
M0	156145.167				
M1	148206.787	M1-M0	7938.380	14.000	<0.0001
M2	146734.413	M2-M1	1472.374	10.000	<0.0001
M3	145973.716	M3-M2	760.697	12.000	<0.0001

Table 5 summarizes results from LR tests. Results from model comparison suggest M3 as the final model. For the sake of brevity, we will discuss herein only results from M3 (Table 6).

Table 6. Final model parameter estimates

	Ability			Speed		
	Estimate	**S.E.**	***p*-value**	**Estimate**	**S.E.**	***p*-value**
Intercept	0.520	0.050	0.000	-0.330	0.069	0.000
male	0.110	0.008	0.000	0.100	0.009	0.000
student's ESCS	0.002	0.004	0.708	0.018	0.005	0.000
student_repeating_one_or_more_grades	-0.149	0.011	0.000	0.225	0.012	0.000
low prior achievement vs average and high	-0.442	0.008	0.000	0.248	0.010	0.000
math test anxiety	-0.162	0.004	0.000	-0.030	0.005	0.000
second generation immigrant vs native	-0.085	0.016	0.000	0.010	0.018	0.593
first_generation_immigrant vs native	-0.090	0.016	0.000	-0.052	0.019	0.006
Class % of stud. with low prior achievement	-0.007	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	0.211	0.029	0.000	-0.164	0.040	0.000
Class % of students repeating grades	-0.001	0.001	0.203	0.003	0.001	0.008
Class average math test anxiety	-0.046	0.026	0.075	-0.289	0.035	0.000
North West vs Center	0.210	0.028	0.000	-0.169	0.039	0.000
North East vs Center	0.251	0.028	0.000	-0.233	0.039	0.000
South vs Center	-0.259	0.027	0.000	0.159	0.038	0.000
South Islands vs Center	-0.504	0.034	0.000	0.356	0.047	0.000
Liceum vs Vocational	0.106	0.038	0.005	-0.251	0.052	0.000
Technical Inst vs Vocational	0.177	0.027	0.000	-0.371	0.037	0.000
Between-class cov. Matrix						
Variance	0.143	0.005		0.289	0.010	
Covariance (ability / speed)	-0.147	0.006				
Within-class cov. Matrix						
Variance	0.401	0.003		0.529	0.004	
Covariance (ability / speed)	-0.225	0.003				

Ceteris paribus, students with low prior achievement are less accurate and spend less time on mathematics items than their peers. A similar pattern of results emerged for the fixed effect of being a student who repeated one or more grades. As for gender, the unique associations with speed and ability are both positive and very similar in size: males are slightly more accurate and work slightly faster than females. Native students outperform students with an immigrant background in ability, and first-generation immigrants work slightly, albeit significantly, slower than the natives. The unique effect of students' ESCS on ability was not statistically significant, whilst a weak, albeit significant, positive effect emerged with speed. Students' self-reported anxiety before and during

the test is negatively related to ability and speed.

After controlling for relevant individual-level predictors, the contextual effect of class ESCS on ability and speed is significant: students from classes with higher ESCS spend more time on items and obtain better results in terms of ability. The percentage of students with an immigrant background is 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 spend more time on items.

Significant differences in ability and speed also emerged by school tracks and geographical area. Students from the vocational school were less accurate and spend less time on the items than those from the lyceum and technical institute. Students from the North-East and the North-West are more accurate and work slowly on items than those from the Center of Italy, whilst those from the South and the South and Islands were less accurate and spend less time on items.

4. Concluding remarks

The main results show that the ability and speed are inversely proportional, e.g. as ability increases, speed decreases. Also, differences in the students' performance by prior achievement, math test anxiety, sociodemographic characteristics, class compositional variables, school tracks and geographical area are significant for both ability and speed. The various results in this study need to be confirmed through additional research. Some further developments should also focus on the opportunity to include response information in the detection of aberrant response behaviour.

References

Charlton, C., Rasbash, J., Browne, W.J., Healy, M., Cameron, B. (2020). *MLwiN Version 3.05*. Centre for Multilevel Modelling, University of Bristol.

Fox, J. P., Klotzke, K., & Simsek, A. S. (2021). LNIRT: An R Package for Joint Modeling of Response Accuracy and Times. *arXiv preprint arXiv:2106.10144*.

Klein Entink, R. H., Fox, J.-P., van der Linden, W. J. (2009). A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers. *Psychometrika*, **74**(1), pp. 21-48.

Rasbash, J., Steele, F., Browne, W.J., Goldstein, H. (2017). *A User's Guide to MLwiN, v3.00*. Centre for Multilevel Modelling, University of Bristol.

van der Linden, W. J. (2007). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, **72**(3), 287.