Education and Science

2025 Supplement 1 129-146

Original Article

How Does Incorporating the Response Times into Mixture Modelling Influence the Identification of Latent Classes for Mathematics Literacy Framework in PISA 2022? *

Halime Yıldırım Hoş [1], Menekşe Uysal Saraç [2]

Abstract

Based on PISA 2022 mathematics literacy test data for Türkiye, this study employed a mixture item response model to identify the ability-and non-ability latent classes of students. In line with the mixture item response theory modelling approach proposed by Jeon and De Boeck (2019), the relations between response times and item difficulty and success probabilities were examined by using four different models in a hierarchical comparison. The first of these models was a single-class two-parameter item response theory (2PL IRT) model (Model I), and the second one (Model II) was a two-class model called the ability class and the guessing class with a success probability fixed at 0.25. In the other two-class model (Model III), the success probability of the guessing class was freely estimated. The final model was a two-class model (Model IV) that included the ability class and the non-ability class, i.e. the one with the response time information as a covariate, in line with the approach proposed by Jeon and De Boeck (2019). As a result of the analysis, Model IV (a two-class model in which response time was included as a covariate) was found to be the best fitting model. Whereas the average item response times and success probabilities tended to be low in the non-ability class, these values were higher in the ability class. However, the ability class, which utilized time more effectively (with higher probability of success), was successful by responding rapidly to easy items while spending more time on difficult ones. As opposed to that, the overall low performance of the non-ability class was noteworthy since it turned out that their faster responses on easy items resulted in failure, whereas they were partially successful by dedicating more time to difficult ones. The latter group seems to have adopted a more superficial approach in which they used a type of item response strategy so that they could respond faster than the ability class on all items but tended to be careful by spending more time on difficult items.

Keywords

Item response time
Knowledge retrieval strategy
Rapid guessing
Mixture item response theory
PISA

Article Info

Received: 10.06.2024
Accepted: 01.24.2025
Published Online: 03.03.2025

DOI: 10.15390/EB.2025.14125

* A part of this study was presented at the International Symposium on Measurement, Selection and Placement held between 4-6 October 2024 as an oral presentation.

[1] İstanbul Medeniyet University, Faculty of Educational Sciences, Department of Educational Sciences, Türkiye, halime.yldrm@gmail.com

[2] Çankırı Karatekin University, Faculty of Humanities and Social Sciences, Department of Educational Sciences, Türkiye, menekseysl@gmail.com

Introduction

In cognitive assessments, typically employed to measure ability, the actual measurements are levels of performance, and processes consist of the actions taken to reach that performance result (De Boeck & Jeon, 2019). Especially small-scale assessments often focus on measures of ability but do not provide measures of item response time. Ability can be assessed without gathering data on processes, but this merely reflects the performance level, and although performance is significant, obtaining further information on response processes (such as item response strategies and times) offers more insight than just showing performance levels. Such data can help to comprehend how students reason and what strategies they employ in addition to offering further information on response processes that give us deeper insight and a narrative of how events unfold. Thus, it not only helps us to understand student response strategies but also gives us an idea about interventions and improvements.

Log data or process data are becoming increasingly common to automatically collect data on the behavior of individuals taking computer-based, large-scale international examinations such as PISA and TIMSS (Anghel, Khorramdel, & von Davier, 2024). Such data include the time it takes test takers to respond to the test or its items, what they click on, their typing order, etc. When based on an existing theory of the cognitive processes underlying people's approach to a test (Brückner & Pellegrino, 2017), such information can help to improve item design, determine test taker engagement, and make inferences about construct (Oranje, Gorin, Jia, Kerr, Ercikan, & Pellegrino, 2017). In this way, log data can provide evidence of assessment validity (AERA, APA, & NCME, 2014). It is also likely to help understand and reframe the differences in achievement in light of different test-taking strategies (Pohl, Ulitzsch, & von Davier, 2021).

Numerous factors can influence the precise evaluation of individuals' abilities in assessment procedures. According to Erwin and Wise (2002), for example, low effort is the most obvious obstacle to the accurate estimation of an individual's ability. When respondents do not end up receiving a grade etc. at the end of an exam, their lack of effort during the exam is perceived as a threat to its validity (Baumert & Demmrich, 2001; Eklöf, 2010; Finn, 2015; Wise, Pastor, & Kong, 2009). After all, to perform well on a test, the respondent definitely needs sufficient knowledge and skills and enough motivation to actively participate and engage in the test (Eklöf, 2010). According to Haladyna and Downing (2004), the fact that the ability estimation of the one with lower motivation between two students at the same proficiency level is lower than the other reflects the difference between their motivation, not the difference in proficiency between these two students. The exam motivation mentioned here is actually a special case of achievement motivation and is the motivation of the examinee to perform well (Baumert & Demmrich, 2001; Eklöf, 2010). Test-taking motivation refers to engage with test items, to exert effort and to remain determined while solving them. This motivation emerges in cognitive strategies such as the effort respondents make to complete the test and the item response strategies they use (Baumert & Demmrich, 2001; Brophy & Ames, 2005). In PISA or other similar assessments, students are usually invited to take a test designed to measure their proficiency in math, reading and science. Students' participation is not mandatory, and their results have no direct impact on them: the test is therefore a low-stakes assessment at participant level (Baumert & Demmrich, 2001; Finn, 2015). Since it lacks grade or personalized feedback for students, low motivation is anticipated when it comes to taking the test. Thus, it is of great importance to make inferences about student effort and motivation, as well as the response strategy used in this exam that offers process data.

The different item-response strategies of test takers have long been of interest to educational researchers. For example, Mislevy and Verhelst (1990) proposed a model where individuals utilize one of a finite number of independent item-response strategies during an exam, and that strategy can be estimated according to response patterns. Yamamoto (1989) suggested a HYBRID model with finite mixture that can also be used to discriminate individuals' multiple item-solution strategies. The researchers paid particular attention to the guessing strategy of test takers, which often occurs during tests administered under time constraints (Jeon & De Boeck, 2019). Mislevy and Verhelst (1990) and Yamamoto (1997) then utilized their models to explore the random guessing item-response strategy.

Moreover, a number of researchers have presented various modelling approaches to capture the guessing strategies of participants under time pressure in speeded tests (Bolt, Cohen, & Wollack, 2002; Cao & Stokes, 2008; Chang, Tsai, & Hsu, 2014; Wang & Xu, 2015). In low-stakes assessments, unmotivated test takers may also use a guessing strategy (Pokropek, 2016; Wise & Kong, 2005; Wise & DeMars, 2006). In the given studies, guessing is usually defined as an item response strategy that does not rely on one's ability. Since guessing is not based on a solution strategy in those models, it is usually assumed that it is a fast process, that its accuracy is at or below the expected accuracy by chance, and that there is only one type of guessing strategy that can be distinguished from a normal item solution strategy (Jeon & De Boeck, 2019). The basic idea underlying rapid guessing is that test takers want to dismiss a given item by selecting a random answer quickly. On the other hand, test takers may also show a solution behavior by using their knowledge, skills and abilities to read the item, figure out its difficulty, and submit a response. Conceptually, solution-oriented behavior requires more time than rapid guessing behavior (Wise, 2019).

Response strategies can be viewed as characterized not only by the patterns of item responses in terms of their accuracy, but also on the basis of response time (Jeon & De Boeck, 2019). In large-scale and low-stakes assessments, test-taker disengagement poses a threat to validity, including the possibility that reduced test participation may result in reduced performance over the duration of the test. This has led to a great deal of interest in 'disengaged rapid guessing' in research (Maddox, 2023). For instance, considering their item responses and times, a group with a low probability of achievement may be labelled as having a 'rapid guessing strategy' if they respond rapidly or in a short time. However, responding rapidly to a difficult item may also indicate one's motivation to avoid a negative experience, a desire to protect one's self-esteem by avoiding an unpleasant situation, a lack of effort, or a shallow approach to coping with tests (Sideridis, Tsaousis, & Al-Harbi, 2022). This means that response times are an informative covariate or predictor in the estimation of the class with a 'guessing strategy'. However, as Jeon and De Boeck (2019) suggested, when response times are used as a covariate, this 'rapid guessing' class may not necessarily be the only alternative to the regular class - the ability class - in which ability estimated.

In an alternative strategy, participants would respond rapidly and accurately to easy items but would spend more time on more difficult items that require both knowledge and reasoning. This strategy is called 'knowledge retrieval' (Jeon & De Boeck, 2019; Sideridis et al., 2022). The knowledge retrieval strategy will require relatively short response times and high probabilities of achievement, especially for easier items, as opposed to relatively more time and lower achievement rates for more difficult items, since getting the correct answer requires reasoning processes beyond knowledge retrieval. In Bloom's taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), knowledge is considered the lowest level of ability or skill development; thus, subjects at more advanced stages of development can be expected to apply to more complex strategies than simply retrieving ready-made knowledge. Irrespective of its source, such information may be present on the surface of the knowledge base; hence, a strategy based on knowledge will be likely to yield a rapid and accurate response. In other cases, a response needs to be constructed based on inferences one makes from other pieces of information, or incorrect responses need to be eliminated, which all require a deeper type of processing, for the correct answer is not available on the surface of a person's knowledge base. The terms 'surface' and 'deep' are associated with learning and comprehension in the literature (Entwistle & Peterson, 2004; De Jong & Ferguson-Hessler, 1996). Another concept related to knowledge retrieval is automatic and controlled processing (Shiffrin & Schneider, 1977). While the retrieval of ready knowledge is consistent with automatic processing, controlled processing is deliberate, requiring effort. Controlled processing requires more time to make inferences or practices. A knowledge retrieval strategy can, in principle, be used with deep as well as surface knowledge, though in the present study the interpretations were based more on automatic processing rather than on controlled processing.

Knowledge retrieval as a strategy for solving questions typically takes place when an individual does not employ their regular ability, when they do not necessarily apply fast response time, or when the accuracy rate may be higher than the expected level due to chance (Jeon & De Boeck, 2019). Guessing can be defined as an item solving strategy that is not based on the individual's ability. The following assumptions were likewise mentioned in previous studies (Bolt et al., 2002; Chang et al., 2014; Wise & DeMars, 2006; Wise & Kong, 2005): (1) Guessing is usually a fast process since a relatively shorter time is spent compared to the ability solution strategy; (2) Correct response rate in the guessing strategy typically aligns with or falls below the anticipated accuracy rate, depending on probabilities; (3) there is single form of guessing strategy that can be differentiated from a typical item-solving strategy.

Consequently, understanding and improving academic achievement requires analyzing and identifying which type of item-response strategy individuals use for test-taking, i.e. during an exam (Jeon & De Boeck, 2019; Sideridis et al., 2022). One important factor is to understand the role of the time taken to respond to the item. The digital transformation of educational testing has provided many new opportunities for technology to enhance large-scale assessments. These encompass the potential to consistently and extensively collect and use log data regarding test takers' response processes. Process data has long been recognized as an important source of validity for assessments. They are now used for multiple purposes throughout the assessment cycle (Maddox, 2023). While the time spent by the respondents on the items can be a marker of many situations such as effort, motivation to take the test, and item response strategy, it is important to consider item difficulty when making comments or analyses.

Including Response Times into Mixture Modelling

This study aims to interpret student effort through response times by means of latent classes, determined by including individuals' response time (RT) data into a mixture model. This model, developed based on Jeon and De Boeck's (2019) study, combines student responses and response times. For a given group (latent class), by applying the 2PL model, the probability that person j responds correctly to item i, Yij, is estimated as follows:

$$P(Y_{ij}=1|a_i,\, b_i,\, \theta_j,\, R_g) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$$

In the given equation, the probability of being successful on a dichotomously scored item is a function of the parameter a, which is the discrimination parameter, b denoting item difficulty, and θ - latent ability estimation. Rg indicates the regular group. In the Mplus software, the discrimination and item difficulty parameters are determined using an item factor analysis (IFA), which is converted into item discrimination using factor loadings:

$$\alpha i = \lambda i\sqrt{f_{var}}$$

In the equation, λ_i indicates the item factor loading and $\sqrt{f_{var}}$ denotes the variance of the latent factor. When the variance of the latent factor is fixed to 1 for identification purposes, the factor loading is equal to the item discrimination. Threshold values are used to calculate item difficulty:

$$\beta_i = \frac{\tau_i}{\lambda i\sqrt{f_{var}}}$$

In the equation given above, τi denotes the threshold estimation in the 2PL model. As in the previous equation, λ_i is the item factor loading and $\sqrt{f_{var}}$ is the variance of the latent factor.

The secondary class S_g is a non-ability class in which the ability is not used in the estimation. The estimation of parameters for this class is carried out in the following manner:

$$P(Y_{ij}=1|\delta_i , S_g) = \frac{e^{(-\delta_i)}}{1 + e^{(-\delta_i)}}$$

In the given equation, δ_i denotes the intercept parameter for the secondary class and is the negative log value of the relevant parameter. Thus, the probability of achievement of person j in the S_g (secondary) class is a function of item difficulty δ_i, but the ability parameter (θ_j) is not included in the probability function in this equation. The probability of an individual being assigned to the secondary class, i.e. non-ability latent class, is estimated using a multinomial regression model as follows:

$$\mathrm{P(Sg)}=\frac{\mathrm{e}^{(\gamma_0\text{-}\sum_i^I \gamma_1 * \mathrm{RT}_{ij})}}{\sum_{u=1}^{1+s} \mathrm{e}^{(\gamma_0\text{-}\sum_i^I \gamma_1 * \mathrm{RT}_{ij})}}$$

In the above-given equation, γ_0 and γ_1 denote constant and slope parameters. The parameter that should be noted here is the parameter γ_i, which shows whether the response times contribute to the assignment to the Sg class. In other words, it helps us to interpret whether spending more/less time on the question provides important information for assigning the individual to the Sg class. If the γ_i coefficients are negative for all items, this may be indicative of a rapid-guessing or fast-responder group. On the other hand, positive γ_i coefficients indicate that spending more time on an item increases the likelihood of being assigned to the Sg group, and if this occurs across all items, it is likely to indicate a thoughtful and careful grouping (as we do not know the qualitative aspects of time). Given that the assignment to the Sg group is based solely on the response times and is independent of ability, the reason why these latent groups are called 'non-ability' groups is not that they lack abilities, but that the ability parameter is not used in their assignment to the specific latent class (Sg). The term non-ability/secondary class is used to refer to a class of respondents who work as not expected on test items designed to measure a specific type of ability (Jeon & De Boeck, 2019). In standard item-response analysis, all test takers are expected to respond to test items based on their ability. In this sense, a class of participants who do not rely on their ability in solving test items is regarded as a 'secondary' class as opposed to the 'usual/regular' ability class (which is assumed and analyzed in the present study).

This study aimed to identify the ability and non-ability latent classes of a group of Turkish students taking the PISA 2022 mathematics literacy test, using the mixture modelling approach proposed by Jeon and De Boeck (2019). Furthermore, it also aimed to contribute to modelling studies in this field by examining how individuals can be classified according to their response times, which is a non-ability variable. In the model constructed in this study, the positive correlation between item difficulty and response time slopes manifests a latent class using the knowledge retrieval strategy. On the other hand, the class characterized by low probability of achievement and rapid item responses may represent a latent group using the rapid guessing strategy. Nevertheless, the model results may reveal different latent classes other than the expected classes, as reported by Jeon and De Boeck (2019) and Sideridis et al. (2022).

Method

Research Model

This study aims to classify the students participated in the mathematical literacy test in PISA 2022 Turkey application by modeling them as regular and secondary classes depending on their item solving strategies. So, it is a descriptive study with respect to determining the latent classes specificized based on the modeling framework that best fits the data and examining the relationships between response times and item difficulties.

Participants

This study focuses on the Turkish data from PISA 2022. PISA targets 15-year-old students worldwide who have reached the end of compulsory education and continue formal education. The stratified random sampling method is used to form the sample of PISA and the strata to be included in the sample are decided jointly by the International Center and the countries. In PISA, a two-stage sampling process is applied. In the first stage, the International Center randomly selects participating schools for each country. In the second stage, the participating countries randomly select the students to be assessed with the help of a computer program called Maple (Ministry of National Education, 2022). The sample of this study consists of 1180 students who took the 13th, 14th, 20th and 24th booklets, which were answered by the highest number of participants in Türkiye in PISA 2022.

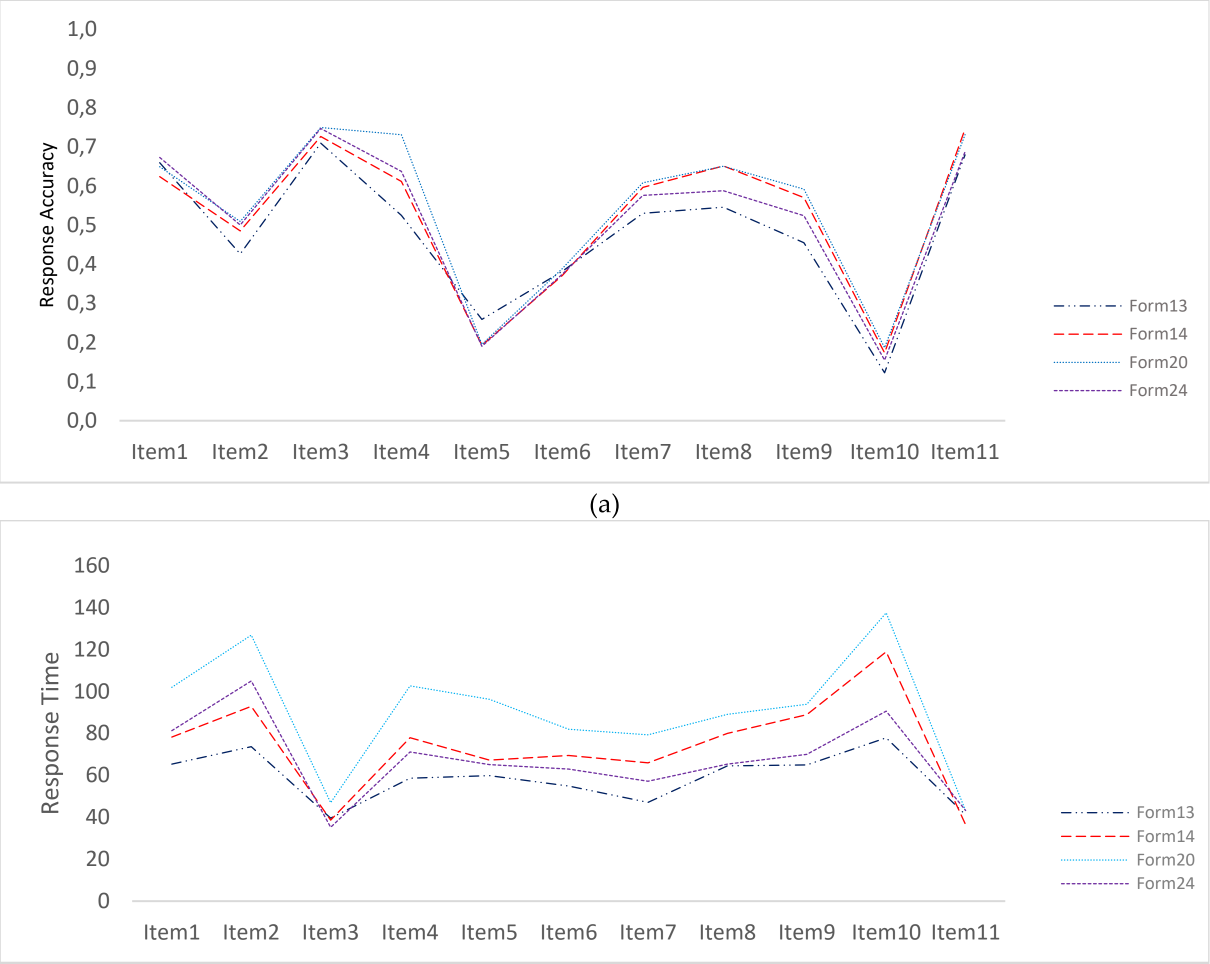
Data Acquisition and Measurement Tools

PISA defines mathematical literacy as an individual's ability to solve real-life problems using mathematical thinking skills. The data containing PISA 2022 cognitive items are open access and accessible through the OECD database. The data used in the study were downloaded from the following link: https://www.oecd.org/en/data/datasets/pisa-2022-database.html. The test items were a mix of easy and difficult items ranging from knowledge to application level.

Data Analysis

When the multiple-choice items in the study were analyzed in terms of response accuracy and response times, response time across test items ranged from 40.03 to 106.049 (seconds) (mean 72.42 and median 63.93). The frequency distribution of item response times is given in Appendix 1. The average correct response rate for all test items was approximately 51.5%, ranging from 2% to 73%, with most test takers having responded to all test items. The rate of omitted responses was very low, and for 90% of the items in the dataset, the rate of inaccessible items was less than 1%.

Figure 1 shows the distribution of response times and response accuracies of the items in the forms included in the analysis. Figure 1(a) further illustrates that the response accuracies do not exhibit a clear pattern according to the item's position in the test. Figure 1(b) demonstrates that the response time distributions are similar to each other although the items are in different positions. The response time distributions do not exhibit a bimodal structure, which clearly indicates the existence of a response class with rapid responses. This suggests that some of the responses may have been guesses, but a significant proportion were not (Jeon & De Boeck, 2019).



(a)

(b)

Note: The item rankings in these images are the result of data labelling only and do not represent the actual layout of the booklets.

Figure 1. Distribution of Response Times and Response Accuracy of the Items in the Booklets

Prior to the data analysis, log transformations of the raw response times were applied to eliminate the skewness in the distributions. The median/mean of the obtained log-transformed response time distribution was found as 4.10/4.02, along with the maximum value of 8.20. In addition, within-person and within-item centering were applied for the log-transformed response times. The aim of double-centered approach was to reduce variations due to time intensity and individual speed differences between items. Double-centered response times were calculated with the following formula: $RT_{ip}^{DC} = RT_{ip} - \overline{RT}_{.p}^{WP} - \overline{RT}_{i.}^{WI} + \overline{RT}_{..}$ (Jeon & De Boeck, 2019).

For the purpose of this study, four models were compared hierarchically. First, the analysis was based on a single class two-parameter item response theory model (Model I). The second model (Model II) consisted of ability and non-ability (secondary) classes. The non-ability class in Model II is a random guessing group with success probability fixed at 0.25, which is the inverse of the number of response options. The other two-class model (Model III) consisted of ability and non-ability classes in which thresholds were freely estimated in both classes. Finally, following the approach proposed by Jeon and De Boeck (2019), the fourth model (Model IV) is a two-class model, consisting of an ability class, and a non-ability class in which response time was incorporated as a covariate.

The given models were evaluated by using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) (Schwarz, 1978), Sample size–Adjusted Bayesian Information Criterion (SABIC) and Scaled Likelihood Ratio Test (SLRTS) (Lo, Mendell, & Rubin, 2001); and the calculations in the models were conducted with Mplus.

Results

Firstly, the data were analyzed and examined according to Model I, Model II, Model III and Model IV, in that order. The model fit indices are presented in Table 1.

Table 1. Fit Indices for Estimated Models

Model	LL	N.Par	c	M-Comp	-2*LL	SLRTS	AIC	BIC	SABIC
Model I: Single class	7120.087	22	1.010	-	-	-	14284.1	14395.7	14325.8
Model II: Two-class (fixed)	7120.091	23	0.967	-	-	-	14286.1	14402.8	14329.7
Model III: Two-class (free)	7093.054	34	1.010	M3-M2	54.074	49.038*	14254.1	14426.5	14318.5
Model IV: With two-class response times	6748.443	44	1.075	M4-M3	689.22	533.06*	13584.8	13806.9	13667.1

Note: p < 0.001 significant. LL = Likelihood Ratio, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion, SABIC = An Adjusted for Sample Size - BIC, SLRTS = Scaled Likelihood Ratio Test Statistics, c = scaling correction factor

The analysis of model fit statistics reveals that Model III (two-class model with free parameters) provides a better fit compared to Model II with LL = 7093.054. In addition, the decrease in AIC = 14254.1 and BIC = 14426.5 values show that this model is more compatible with the data than the fixed two-class model. The SLRTS test results show that the difference between the two models is statistically significant with χ^2 (11) = 49.04, p < .001. With an LL value of 6748.443, the two-class model that incorporates response times offers the best fit in the final phase. The significant decreases in AIC (13584.8), BIC (13806.9), and SABIC (13667.1) values in this model indicate a notable improvement in fit for this model. Compared to Model III, the SLRTS test results show a significant improvement with χ^2 (10) = 533.06, p <.001. In conclusion, Model IV performs the best across all fit criteria, and including response times in the analysis greatly improves the model's fit to the data.

The Two-Class Model with Response Times Included as a Covariate

Table 2 presents the average values of the posterior probabilities and the entropy value for Model IV, which includes two classes, with response times as a covariate (ability-usual R_g and non-ability-based-secondary S_g class) model.

Table 2. Average Values of Posterior Probabilities and Entropy Value for Model IV

Latent Classes	Average Values of Posterior Probabilities (AvePP)		Entropy	Membership in the Most Likely Class
	Non-ability class	Ability class		n (%)
Non-ability class	0.943	0.057	0.814	551 (48%)
Ability class	0.049	0.951		597 (52%)

The averages of the posterior probabilities and the entropy value presented in the table indicate that the AvePP is above 0.90, with the entropy value reaching 0.814. This indicates that the latent classes are strongly separated and that the model has low classification uncertainty. In the potential class distribution, there are 551 people (48%) in Class 1 and 597 people (52%) in Class 2. This balanced distribution indicates that both classes are distinctly defined and that the model is performing well.

Comparative parameter estimates for the ability and non-ability classes related to mathematics literacy items are provided in Table 3.

Table 3. Item Parameter Estimates for Ability and Non-Ability Classes

Mathematics Literacy	**Ability class**				**Non-ability class**			
	Slope $\lambda_i^{R_g}$		**Threshold value** $\tau_i^{R_g}$		**Threshold value** $\delta_i^{S_g}$		**RT Slope** γ_i^{Sg}	
	Est.	**SE**	**Est.**	**SE**	**Est.**	**SE**	**Est.**	**SE**
Item1	0.669**	0.243	-1.332**	0.118	-0.002	0.098	-0.011	0.249
Item2	0.989**	0.286	-0.866**	0.122	1.205**	0.121	-1.404**	0.189
Item3	0.738*	0.300	-1.984**	0.16	0.691**	0.119	-0.697**	0.206
Item4	-0.054	0.261	-1.798**	0.131	-0.315**	0.099	1.734**	0.288
Item5	0.229	0.188	1.262**	0.111	1.389**	0.121	-0.055	0.242
Item6	0.395	0.227	-1.277**	0.108	0.686**	0.104	-0.481*	0.237
Item7	0.464**	0.163	0.189	0.092	0.889**	0.106	-1.008**	0.239
Item8	1.142*	0.498	-2.562**	0.287	1.036**	0.127	0.807**	0.218
Item9	0.452	0.271	-1.717**	0.134	1.675**	0.165	-1.745**	0.256
Item10	1.545**	0.483	1.297**	0.232	6.551**	2.232	-1.624**	0.23
Item11	0.799	0.374	-2.528**	0.225	0.146	0.105	-0.398*	0.152

Note: RT = Response Time. *p<.05, **p<.001.

The estimated threshold parameters for the ability class indicate that the threshold values for most of the items (except for Item 5, Item 7, and Item 10) are negative, revealing that participants in the ability class generally found the items easy. On the other hand, threshold parameters estimated for the non-ability class show that most items (except for Items 1 and 11, which are not significant, and Item 4, which is negative) have positive values, indicating that participants in the non-ability class found many items difficult. For example, the threshold value of Item 1 (-0.002) indicates that this item is perceived as having average difficulty by the participants in the non-ability group, whereas items with high positive threshold values like Item 2 (1.205), Item 9 (1.675), and especially Item 10 (6.551) were quite challenging in the non-ability class.

It is also noteworthy that there are negative slope coefficients for item response times in the non-ability class, which indicates that participants who spend more time on the relevant items reduce the likelihood of being assigned to the non-ability class. For example, if a student spends more time on Item 10, the likelihood of being in the secondary class or the non-ability class is reduced. Upon closer examination, Item 10 stands out with a high threshold value and a negative RT slope coefficient, revealing that students found this item considerably difficult. However, responding to it rapidly (in a shorter time) increased the likelihood of being placed in the non-ability class. Likewise, for Item 9, characterized by a high threshold value and a negative RT slope coefficient (γ_i), it can be concluded that students in the non-ability class encounter difficulties with this item, and as the time spent in the item decreases, the probability of belonging to the non-ability class rises.

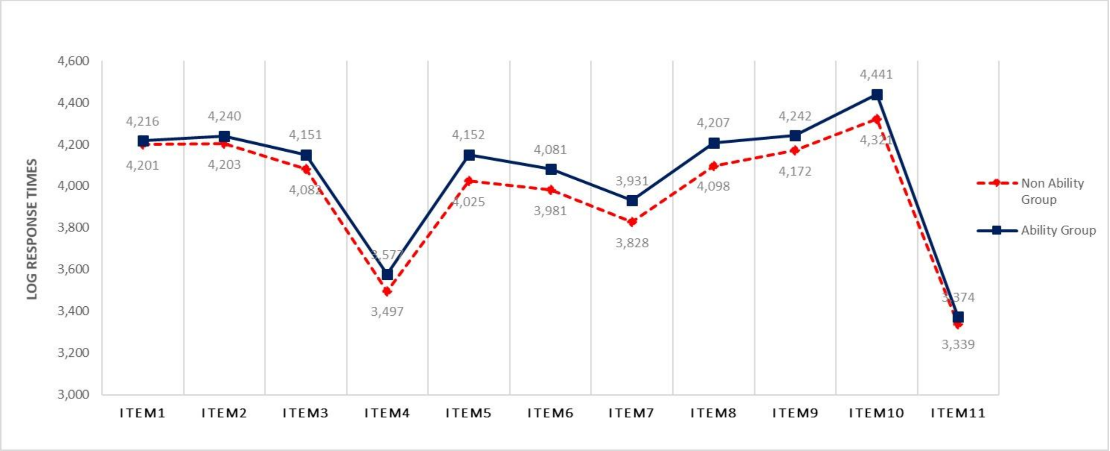
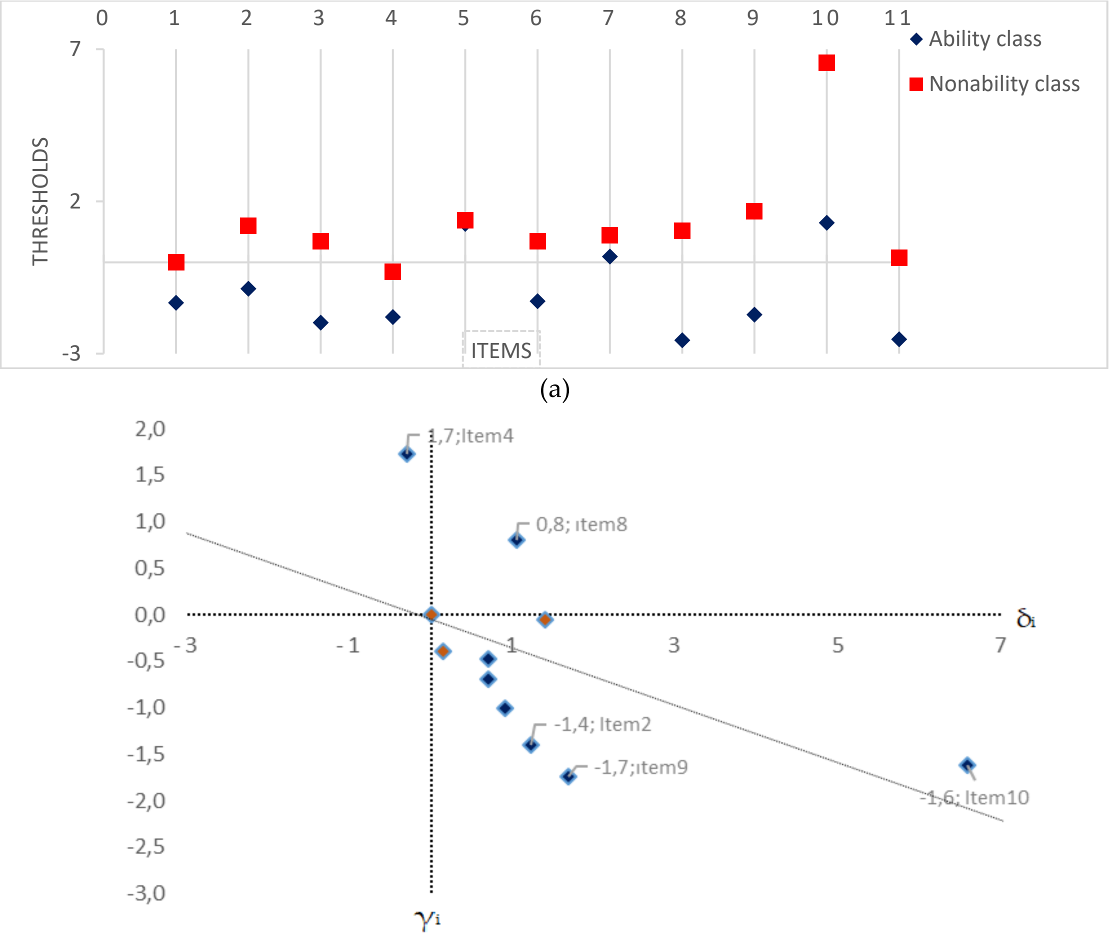


Figure 2. Average Log Response Times

Figure 2 provides the average log response times for the ability and non-ability latent classes. The response time distributions for the two latent classes show that, for all items, the average log item response times of the ability class are higher. While there is no substantial difference in response times, it is important to note that those in the non-ability group exhibited lower average log item response times for all items, suggesting that despite the similar patterns observed in both groups, the ability class took more time to solve the items. Conversely, more rapid response times for the non-ability group might suggest that this group demonstrates reduced cognitive involvement or relies on guessing-based response strategies.



(a)

(b)

Figure 3. (a) Distribution of Threshold Parameters for Ability and Non-Ability Classes and **(b)** Threshold Parameters (δi) and Response Time Slopes (γi) for Items in the Non-Ability Class

The threshold values for the items are provided for the ability and non-ability classes (see Figure 3(a)). The threshold values for the ability class largely show a distribution on the negative axis side, while those of the non-ability group are different on the positive axis side with much higher values in some items (especially Item 10). It is obvious that the non-ability class experienced a distinct difficulty in certain items (Items 2, 3, 10) compared to the ability group; however, in other items (Items 5, 7), they experienced a difficulty closer to that of the ability group, revealing how the difficulty differences between the two latent groups changed on a per-item basis.

As can be seen from the scatter plot showing the relationship between item response times' slope coefficients and item threshold parameters for the non-ability class (Figure 3(b)), what stands out is the negative slope coefficients corresponding to high threshold values in the positive axis direction of most of the items. A negative item response time slope coefficient was determined especially for Item 10 (the rightmost point), which has the highest threshold parameter at the item level. Despite having the highest threshold value, the negative response time slope coefficient of Item 10 can be interpreted as indicating that students are more likely to be placed in the non-ability class when they spend less time solving this item.

There is a significant negative correlation between the item response times and the slope parameters of the non-ability class (Spearman rho= -0.682, p<.05). This result clearly indicates that the group with a rapid response does not employ a knowledge retrieval strategy. As noted by Jeon and De Boeck (2019), "Hence, we expect a positive relationship between item difficulty and the effect of response time on the marginal probability of belonging to the knowledge retrieval class." (p.698)

This study did not name this class as the rapid-guessing group in the non-ability class, as it generally/adoptively embraced the fast response strategy, with no observed negative slope coefficients for all items (Items 4 and 8 had a positive slope).

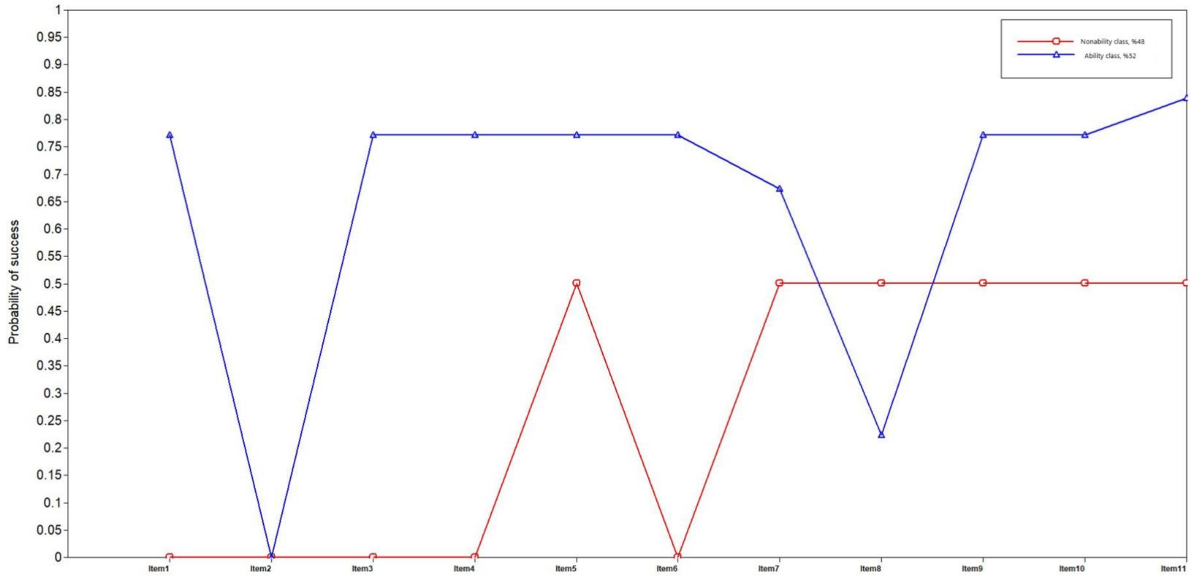


Figure 4. Probability of Succes on Items for Participants in Ability and Non-Ability Classes

As can be seen from the success probabilities of the ability and non-ability classes, the probability of correct responses in the former class generally shows a higher (except for Item 2 and Item 8). For most of the items, the likelihood of the ability class responding accurately lies between 65% and 85%. Particularly for items 1, 3, 4, 5, 6, 7, 9, 10, and 11, the probability of success is heightened; this suggests that the ability class is more competent in certain mathematical literacy topics and tends to respond better to these items.

It has been noted that the ability group's response times are correlated with the items' difficulty. This group performed particularly well by spending less time on the easy items (Items 4 and 11).

Likewise, they succeeded by devoting more time to the most challenging items (Items 5 and 10). The interactive relationship between response time and item difficulty is demonstrated by the fact that the students with high probability of achievement, responded to more slowly to difficult items and more rapidly to easy ones (see Figure 3(a) and Figure 4). It might also point to a latent class that makes good use of information and manages exam time. It is interesting to note that the students spent less time on item 7, thereby decreasing their probability of success, even if they found it challenging.

The success probabilities of the non-ability class generally remained low; in many items, the probability of correct responses was less than 20%. Especially in Items 1, 2, 3, 4, and 6, the probability of success for that group was almost 0%. Interestingly, the non-ability class used a relatively less item response time when responding to Item 4, which is relatively easy (conditionally easy for the class), and their probability of success on this item decreased. Item 3 and Item 6 also exhibited a similar pattern. In these items, which were found to be easier compared to the others, the lowest item response times were spent, yet the success probabilities remained quite low. In the more difficult items (e.g., Items 5, 9, and 10), a rise in response times was noted, alongside an increase in the probability of success. This situation indicates inefficient use of time on relatively easy items, whereas on harder items, it points to a partially successful (probability of success approximately .50) latent class that uses time more effectively.

This detailed examination of item difficulties and response times, when combined with negative response time slopes, suggests the presence of a latent class with a response strategy characterized by "responding rapidly but using time inefficiently." In other words, students may attempt to respond rapidly without thinking on relatively easy items. This may lead to a superficial evaluation, potentially resulting in incorrect answers. In more challenging items, the gradual improvement in students' success with additional time spent might indicate their efforts to engage with the questions with a more in-depth and careful approach.

Discussion, Conclusion and Suggestions

In this study, a mixture modelling method was utilized with PISA 2022 Türkiye data to identify both the ability and non-ability latent classes of students who took the mathematics literacy assessment. Researchers frequently highlight that individuals with low motivation tend to rely on guessing methods, particularly in low-stakes assessments (Wise & Kong, 2005; Wise & DeMars, 2006). This study investigated the presence of latent classes exhibiting varying item response strategies by using the informative aspect of item response times in a low-stakes with no direct consequences for participants, such as PISA.

This study also employed a finite-mixture item response model that accommodates within-person variability in relation to response times. In this model, the secondary class does not have to be categorized as fast or slow, and its nature is not predefined. As emphasized in the researches conducted by Joen and De Boeck (2019), Sideridis and Alahmadi (2022), and Sideridis et al. (2022), the nature of a potential secondary class is derived from data by examining the influence of the characteristics of this class as well as their item success rates, and the covariate of response time. Consequently, the essence of the secondary class is an empirical issue.

This approach differs from the class definitions observed in previous mixed modelling studies; for example, Meyer (2010) established the nature of the classes beforehand in the research. Here, one class is defined as the rapid guessing class, while the other class is identified as the one with solution behavior (ability-based). The classes were separated by the fact that the prior mean in the response time distribution of the solution behavior class was higher than that of the rapid guessing class. Conversely, Wang and Xu (2015) similarly utilized a two-class mixture model where the nature of the classes was predetermined. The authors designated one class as possessing item-specific success probabilities, while the other was deemed as the guessing class. In the current study, classes emerged in a more flexible, data-driven manner, and the characteristics of the secondary class were shaped by data.

To achieve this, the models suggested by Jeon and De Boeck (2019) were evaluated in this study. Specifically, the two-class model that including response times (Model IV) displayed the best fit, and employing response times as a covariate provided a clearer understanding of the differences among the latent classes. The correlations between the parameter estimates for the latent classes derived from Model IV and the item response times in these classes offer significant insights into item response approaches.

In this study, the ability class was defined as a group with higher achievement and longer response times. In particular, the ability class's success in easy items with less time spent may reveal this group's mastery of the subject as well as their quick-thinking skills. Moreover, the fact that they succeeded by spending more time on difficult items suggests an interactive relationship between response time and item difficulty, a finding which may indicate that students with high proficiency levels are effective and strategic in their use of knowledge. However, such interpretations should be evaluated with caution, as they remain speculative.

Furthermore, the secondary, or non-ability class, refers to a group that has lower achievement and uses shorter response times compared to the ability class. However, in the study by Sideridis and Alahmadi (2022), the researchers observed that the latent class defined by response times spent more time and achieved higher levels of achievement than the ability class. The same study also highlighted that additional time may not be as advantageous as it is in other subjects, especially when analytical skills, such as math, are critical to achievement. Moreover, that study discovered that the impact of response time on student performance may differ based on the type of content, suggesting that response times and the complexity of content are substantial factors in assessing student performance. In current study, especially since the accurate response rate for many items remained below 20%, with the success rates being low even for the easiest items, this could indicate that this group lacks sufficient knowledge (even fundamental knowledge) on the subject matter. In difficult items, however, despite spending more time, the partial increase in success rates may reflect that this group put in more effort against harder questions but lacked sufficient knowledge and competence.

Sideridis and Alahmadi (2022) concluded in their study that lower-performing groups spent more time on difficult items but still achieved lower success compared to the highly skilled group. The authors suggested that the additional time invested by low-performing individuals in the items is likely to point to certain content beyond their current skill levels, and it was therefore not beneficial. The study also emphasized that the comments made regarding item-solving strategies were hypothetical/speculative.

In conclusion, this study provides important insights into understanding the relationship between students' knowledge and response strategies, clearly highlighting the necessity of considering these processes in education. Additionally, the Standards for Educational and Psychological Testing, developed by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, recommend the collection and reporting of effort measurements and their use in the interpretation of test scores (International Test Commission, 2013).

Large-scale digital exams and e-exams provide substantial opportunities for data gathering by engaging a broad student population. These applications offer numerous data sources to assess student performance, and the analysis of this data can enhance the understanding of educational systems. In particular, similar to international applications such as PISA and TIMSS, the collection of data on item response time in national-level e-exam applications is critically essential for assessing student achievement and gaining insights into their response strategies. The analysis of such data can enable educators to gain a deeper insight into student achievements (Jeon & De Boeck, 2019).

Assessing student motivation is critically important in establishing whether a test assess knowledge or motivation (Eklöf, 2010). Test performance not only indicates the knowledge levels but also reveals the strategies that students use and motivational states during the test. Digital transformation and technology provide many opportunities to enhance and transform large-scale assessments. Systematic monitoring and analysis of response times are key to these opportunities. In this context, response times can be a powerful tool to improve the quality and reliability of large-scale tests (Maddox, 2023).

In current study, model estimations characterising the non-ability class were limited to two classes. The main reason for this is the difficulties that may be encountered in interpretation and parameter estimation. Similar difficulties were mentioned in Sideridis and Alahmadi (2022), where relevant model comparisons were limited to two classes, and this approach was preferred due to the characterisation of classes and ease of interpretation. Considering the latent nature of the classes, model comparisons with three or more classes can be included, taking into account that there is one ability class but there may be two, three or four non-ability classes. So, it is important for future studies to determine how the response strategies of these classes will change in the presence of more than one non-ability class.

Analyzing response times helps better understand students' strategies. For example, students dedicating more time may indicate knowledge retrieval strategies (Jeon & De Boeck, 2019). On the other hand, shorter response times may also suggest psychological states such as skimming through keywords, avoidance motivation, or learned helplessness (Abramson, Metalsky, & Alloy, 1989; Seligman, 1972). In future studies, the model used in this study can be improved; for example, it may be beneficial to incorporate other behavioral, demographic, or psychological information into the model to investigate the nature of response strategies of classes. The data utilized in the study is large scale yet poses minimal risk to the students. The strategy classes derived from analyzing exams with purposes such as selection, placement, or scoring/grades for students can be re-examined.

References

Abramson, L. Y., Metalsky, G. I., & Alloy, L. B. (1989). Hopelessness depression: a theory-based subtype of depression. *Psychological Review 96*(2), 358-372. doi:10.1037/ 0033-295X.96.2.358

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Anghel, E., Khorramdel, L., & von Davier, M. (2024). The use of process data in large-scale assessments: A literature review. *Large-scale Assessments in Education*, *12*(1), 13. doi:10.1186/s40536-024-00202-1.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: the effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441-462.

Bloom, B., Engelhart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals handbook I: Cognitive domain*. New York: David McKay Company.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331-348. doi:10.1111/j.1745-3984.2002.tb01146.x

Brophy, J., & Ames, C. (2005). *NAEP testing for twelfth graders: Motivational issues*. Washington, DC: National Assessment Governing Board.

Brückner, S., & Pellegrino, J. W. (2017). Contributions of response processes analysis to the validation of an assessment of higher education students' competence in business and economics. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 31-35). New York: Springer International Publishing.

Cao, Y., & Stokes, L. (2008*).* Modeling response times in test-taking: Applications and developments**.** *Journal of Educational Measurement*, *45*(2), 135-153.

Chang, Y. C., Tsai, C. C., & Hsu, H. C. (2014). The impact of guessing strategies on item response theory model parameters. *Educational and Psychological Measurement*, *74*(1), 69-85.

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

De Jong, T., & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist*, *31*(2), 105-113.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education Principles Policy Practice, 17*(4), 345-356. doi:10.1080/0969594X.2010.516569

Entwistle, N., & Peterson, E. (2004). Conceptions of learning and knowledge in higher education: Relationships with study behavior and inferences of learning environments. *International Journal of Educational Research*, 41, 407-428.

Erwin, T. D., & Steven L. W. (2002). A scholar-practitioner model for assessment. In *Building a scholarship of assessment* (pp. 67-81). San Francisco: Jossey-Bass.

Finn, B. (2015). *Measuring motivation in low-stakes assessments (Research report RR-15-19).* Princeton, NJ: Educational Testing Service.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement:Issues and Practice, 23*(1), 17-27. doi:10.1111/j.1745-3992.2004.tb00149.x

ITC. (2013). *International guidelines for test use*. Retrieved from http://www.intestcom.org/Guidelines/Test+Use.php

Jeon, M., & De Boeck, P. (2019). An analysis of an item-response strategy based on knowledge retrieval. *Behavior Research Methods*, 51, 697-719. doi:10.3758/s13428-018-1064-1

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767-778. doi:10.1093/biomet/88.3.767

Maddox, B. (2023). The uses of process data in large-scale educational assessments. *OECD Education Working Papers,* 286, Paris: OECD Publishing. doi:10.1787/5d9009ff-en

Meyer, J. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*(7), 521-538.

Ministry of National Education. (2022). *PISA 2022 international student assessment program.* Retrieved from https://pisa.meb.gov.tr/meb_iys_dosyalar/2022_01/26105818_PISA_2022_TanYtYm_KitapcYYY.pdf

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195-215.

Oranje, A., Gorin, J., Jia, Y., Kerr, D., Ercikan, K., & Pellegrino, J. W. (2017). Collecting, analysing, and interpreting response time, eye tracking and log data. In K. Erickan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 39-51). Mount Royal, NJ: National Council on Measurement in Education.

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

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300-325. doi:10.3102/1076998616636618.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Seligman, M. E. (1972). Learned helplessness. *Annual Review of Medicine,* 23, 407-412. doi:10.1146/annurev.me.23.020172.002203

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127-190.

Sideridis, G., & Alahmadi, M. T. S. (2022). The role of response times on the measurement of mental ability. *Frontiers in Psychology*, 13. doi:10.3389/fpsyg.2022.892317

Sideridis, G., Tsaousis, I., & Al-Harbi, K. (2022). Identifying ability and nonability groups: incorporating response times using mixture modeling. *Educational and Psychological Measurement*, *82*(6), 1087-1106.

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

Wise, S.L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, *32*(4), 325-336, doi:10.1080/08957347.2019.1660350

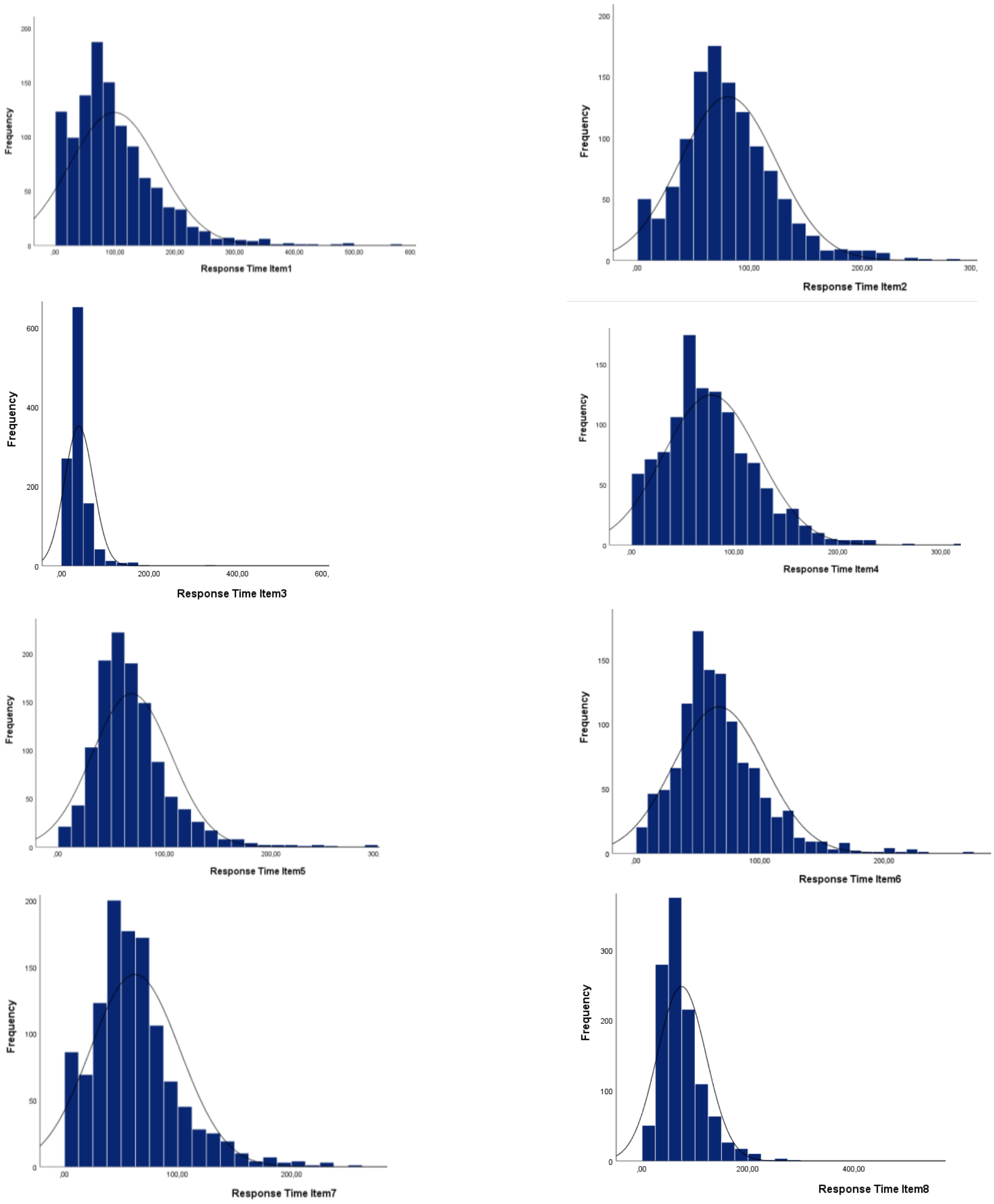
Wise, S. L., & Demars, C. E. (2006). An application of item response time: The effort moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

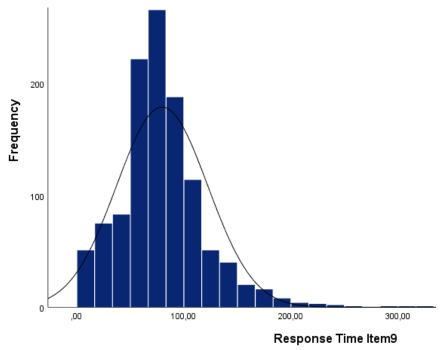
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18, 163-183. doi:10.1207/s15324818ame1802_2

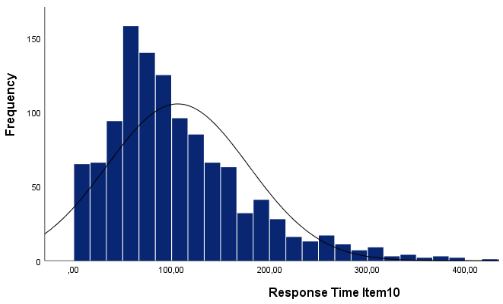
Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185-205.

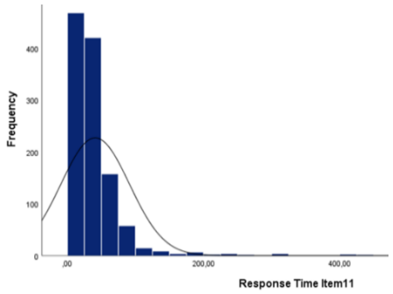
Yamamoto, K. H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). New York: Waxmann Verlag GmbH.

Appendix 1. Distribution of Response Times in Mathematical Literacy Items




Frequency
200
100
0
,00
100,00
200,00
300,00
Response Time Item9


Frequency
150
100
50
0
,00
100,00
200,00
300,00
400,00
Response Time Item10


Frequency
400
300
200
100
0
,00
200,00
400,00
Response Time Item11