Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione

WORKING PAPER N. 50/2020

Estimation of test-taking effort on INVALSI computer-based tests

Chiara Sacco - INVALSI

https://orcid.org/0000-0002-3958-8353

Collana: Working Papers INVALSI

ISSN: 2611 - 5719

Abstract

In the last years, there has been a growing interest in the study of the impact of student disengagement on the test performance and on the validity of the resulting test score in the framework of large scale assessments. The wide diffusion of computer-based tests had opened the way to the development of new measures of motivation, based on the possibility of measuring more aspects of a test than only the student's answer. Thus far, to measure the student's engagement, the most studied variable is the response time at item-level which corresponds to the time spent by each student to respond to a single item. Although the amount of time that a student spends on a single item is a complex phenomenon, affected by several factors as item characteristics, student characteristics and the situation during the test event, the response time has been demonstrated to be a good indicator of the behaviour engaged by the test-taker during the test. This paper investigates the test-taking behaviours using response time collected by INVALSI computer-based tests and presents a new procedure to classify the students' non-effortful responses.

Parole chiave: Tempo di risposta; Impegno; Rapid-guessing behaviour; Effortful responses; Large-scale assessment

Keywords: Response time; Engagement; Rapid-guessing behaviour; Effortful responses; Large-scale assessment

Introduction

In the last years, in the framework of large scale assessments, there has been a growing interest in studying the impact of student's disengagement on the test performances and on the validity of the resulting test score. National and international large scale assessments are widely used to assess and to monitor students' competences (INVALSI, 2019; OECD, 2016) and are used as benchmark from policymakers for planning educational reforms (Breakspear, 2012; Fullan, 2009). The validity of the test score depends partially on the assumption that the students are motivated and engage the test in an effortful manner (Eklof, 2010; Wise, 2015). The impact of disengaged test-takers on score validity has been widely studied (Wise and DeMars, 2005; Braun *et al.* 2011): the students' score is not only the result of what the student knows and can do but also of the effort during the test event, and this threatens the validity of the resulting test score.

Several methods have been proposed in the literature to assess the test-takers engagement. Self-report instruments are widely used to investigate the student engagement but it is well known that these are affected by different types of bias: students could overstate or under-report their effort (Wise and Kong, 2005). Another important limitation is that self-report scales usually provide global information about the student's test-taking motivation during the whole test (Eklof, 2010; Wise, 2014).

The wide diffusion of computer-based tests had opened the way to the development of new measures of motivation, based on the capability of computer-based tests to measure more aspects of a test than only the answer of the student. Thus far, to measure the student's engagement, the most studied variable is the response time (RT) at item-level (Wise, 2017; Wise, 2019), which corresponds to the time spent by each student to respond to a single item. It is reasonable to suppose that very rapid responses are indicative of a non-effortful behaviour of the test-taker (Schipnike, 1995; Wise and Kong 2005). A too rapid response, faster than the time required to read fully, understand and select an answer, suggests that the student did not engage the item in an effortful manner; this behaviour is termed *rapid-guessing behaviour* whereas, when the test-taker engages the item in an effortful manner, the behaviour is termed *solution behaviour* (Schipnike 1995, 1996). An important advantage of this approach is the possibility to study the student's effort at student and item level.

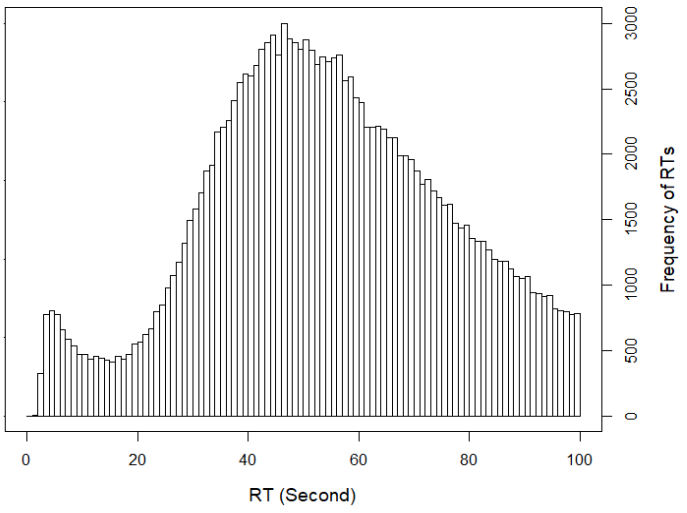
Figure 1 displays the ideal distribution of the response time to identify the *rapid-guessing behaviour* and *solution behaviour*: a bimodal distribution with one mode in the first few seconds corresponding to the rapid-guessing frequency, and another occurring later at the modal response time for *solution behaviour*. In practice, to classify the behaviour of the student for each response as *rapid-guessing behaviour* or *solution behaviour*, it is necessary to identify a RT threshold. Several methods have been proposed in the literature: (a) visual inspection (VI) of the RT distribution based on the conceptual distribution of RT for *rapid-guessing* and *solution behaviour* as explained in Wise (2017); (b) a common k-second method where a fixed threshold,

usually of 3-5 seconds, is defined for all items (Wise et al., 2004); (c) estimation of mixture model to identify the two type of responses (Kong et al., 2007); (d) inspection of both RT and response accuracy distribution (Ma et al., 2011; Guo et al., 2016); (e) percentage of the average RT (normative threshold method, NT10, see Wise and Ma, 2012).

The definition of *rapid-guessing behaviour* can be applied only on the closed-ended items. It has been showed that, in presence of open-ended items and/or when the student can choose to omit the answer, the RT classification based only on the identification of *rapid-guessing* is a weak indicator of effortful behaviour. Wise and Gao (2017) suggested the *rapid-omit* indicator to identify another type of non-effortful behaviour where the examinee viewed the item but left it quickly without answering.

The main aim of this work is to investigate the test-taking behaviours using RT collected from INVALSI computer-based tests. The INVALSI test platform allows the test-taker to answer items in any order he chooses, to review and possibly change answers and to omit them. The identification of the non-effortful response is based on the use of two indicators: the *rapid-guessing* indicator and the *rapid-omit* indicator. Since this is a very first attempt to analyse test-taking behaviours in INVALSI assessments, this work is focused on the identification of the non-effortful responses for only the closed-ended items. The first part of the work put the stress on the identification of the best threshold for each closed-ended item to identify the *rapid-guessing* responses. A new procedure to identify the RT threshold for *rapid-guessing* responses has been proposed and a validation step has been performed to ensure the reasonableness of the resulting indicators. In the second part of the work, the relationship of effortful and non-effortful responses with student and item characteristics has been studied.

Figure 1. Distribution of response time for *rapid-guessing* and *solution behaviours*.



Methods

INVALSI data

National Institute for the Evaluation of the System of Education and Training (INVALSI) conducts every year large-scale survey assessments in Italy to monitor students' skills in Italian Language, Mathematical knowledge and English Language. In particular, INVALSI carries out annually standardized tests to assess the competence of the students at the end of the second and fifth year of the primary school, at the end of the lower secondary school, at the second and at the last year of the higher secondary school. In 2018 INVALSI moved from *paper-and-pencil* to computer-based tests, which tests were administrated by INVALSI for the first time to the students at the 8th grade (last year of the lower secondary school) and at the 10th grade (second year of the higher secondary school). In this work, the mathematics standardized tests administered by INVALSI at the 8th-grade students in the school year 2017-2018 have been exploited. At the 8th grade, INVALSI test is compulsory for the student admission to the state final exam. Moreover, to attest the competence level of the student INVALSI releases a certificate to attest the student's competence.

Twelve test forms were administrated, using a total of 275 mathematics items of which 145 are closed-ended (the test-taker makes a choice between different answer options). The test administration software is developed to guarantee to the test-taker a greater control over test administration. The key idea was that the computer-based test should guarantee the test-taker to have as much control during the test event as the control permitted with the paper version. The platform is designed to allow the test-taker to: answer items in any order he chooses; flag items for possible later review; review and possibly change answers any time he wants; skip items without answering. During May 2018, the computer-based mathematics tests were administrated to more than 500000 students from 5790 schools in Italy. For each item in addition to the student raw responses, the RT, i.e. the total time spent on the item by the student, has been collected.

Defining of RT threshold

The first step for identifying rapid-guessing responses is the definition of a threshold of the RT; if the RT of a single item is equal or below the threshold the student response is classified as *rapid-guessing*, otherwise the item response is classified as *solution behaviour*. The four threshold identification methods were applied to determine the RT threshold for each item:

- Common method (Fix5): where a fixed threshold of 5 seconds is applied to all items;
- Normative threshold (NT10): the RT threshold for each item is set at 10% of the average time spent on the item by all the test-takers, with a maximum threshold value of 10 seconds;

- Response time and accuracy method (RA10, RA20): this method requires to compute the accuracy, i.e. the proportion of the correct responses for each item in correspondence of each time *t*. It is assumed that the accuracy of rapid guesses tends to the random chance level. Focusing on the first 20 (10) seconds the threshold, called RA20 (RA10), is set equal to the last time *t* where the accuracy is below the chance level.

The evaluation of the four different thresholds provides a measure of the reliability of the use of *rapid-guess* and as indicator of non-effort. In the present study, two different criteria are used; item responses correctly classified as *rapid-guess* should:

- Be correct at a rate much lower than those classified as *solution behaviour*
- Be correct at a rate close with the one expected from random responding

Once the two criteria are verified, to choose the RT thresholds the items were classified according to two different characteristics of the item RT distribution: the presence of more than one mode and the skewness in the first 100 seconds. In particular, for each item, the Hartigan's Dip Test is applied to verify the unimodality of RT distribution. A measure of skewness of the RT distribution is computed in correspondence of each item; only the items with skewness lower than -0.3 were classified as left-skewed. As suggested by Lee and Jia (2014) not all the items show a clear bimodal RT distribution (see Figures in Appendix). Since the visual inspection of the items to examine the properties of the RTs is very demanding, the classification of the RT distribution using the Hartigan's Dip Test and the skewness helps in the identification of RT threshold for each kind of pattern: bimodal and left-skewed, bimodal and no left-skewed, no bimodal and left-skewed and not bimodal and no left-skewed. The identification of the best threshold for each group of items depends on the absence of a positive relationship between the mean accuracy and test-taker achievement level.

For the identification of the *rapid-omits* only the NT10 threshold has been used.

Effortful indicator

For each item with an RT threshold identified for the *rapid-guessing* indicator and for the *rapid-omit* indicator, a dichotomous *index of effortful behaviour* (EB) for each student *i* and item *j* has been computed (Wise and Kong, 2005; Wise and Gao, 2017) as follows:

$$EB_{ij} = \begin{cases} 1 & if\ RG_{ij} = 0 \wedge RO_{ij} = 0 \\ 0 & otherwise \end{cases}$$

where RG_{ij} is the *rapid-guessing indicator* and RO_{ij} is the *rapid-omit indicator*. If the index of EB is equal to 1 then the student response at item level is classified as *non-effortful*, otherwise as *effortful*. This index was used as indicator for effortful behaviour at student-item level.

The mean accuracy and mean RT of the two response groups were analysed in order to guarantee the accordance of the final indicator with the two validation criteria. An additional validation criterion was evaluated for assessing the reliability of the proposed EB indicator. The mean accuracy and the mean response time of the two responses groups were studied with respect to two items characteristics: the item dimension (Arguing, Knowing and Problem Solving) and the item position. In this analysis the item position is categorized in 3 classes: 1st block if the item has been administered in the first ten positions, 3rd block if the item has been administered in the last ten positions, 2nd block otherwise. The last validation hypothesis is that non-effortful responses should have an accuracy rate consistent with that expected from random responding regardless of the item position or dimension.

To investigate how students test-taking behaviour is related to students and item characteristics, the following measures were computed:

- *Response behaviour effort* (RBE; Wise and Gao, 2017): the EB_{ij} is aggregated across all the responses made by a test-taker. This student-level measure corresponds to the proportion of items for which the student exhibits an effortful behaviour. A graphical analysis of the distribution of the RBE indicator conditional to a self-reported measure of the student interest for his/her scholastic carrier was performed.
- *Response behaviour fidelity* (RBF; Wise and Gao, 2017); the EB_{ij} is aggregated for a given item across all the students. This item-level measure represents the amount of effort received by the item and is equal to the proportions of students exhibiting effortful behaviour to an item. An analysis to assess the effect of the position, the item difficulty and the item dimension (Knowing, Problem Solving, Arguing) on the RBF was performed. Each item can be administrated in different forms: same difficulty and dimension but a different position. So, to estimate the effect of these item characteristics on the RBF a linear mixed model was computed.

Results

Defining of RT threshold for rapid-guessing indicator

The first validation hypothesis for the *rapid-guessing* indicator is that *rapid-guessing* responses should be correct at a rate much lower than that from solution behaviours. This hypothesis was supported for all four threshold methods by the results shown in Table 1, which reports the overall accuracy rates for the two types of responses for each threshold methods. All thresholds lead to an averaged accuracy of *rapid-guessing* lower than the accuracy of the solution behaviours. In particular, mean accuracy for the *solution behaviours* is found

around 0.50, whereas *rapid-guessing* responses show an accuracy rate that ranges from 0.17 to 0.26. These results verify the second validation hypothesis, too. The second validation hypothesis was that *rapid-guessing* responses should be correct at a rate close with the expected from random responding. The accuracy of *rapid-guessing* responses is always close to the chance level, which can be assumed to range between 0.1 to 0.25 according to the number of choices of the item. While, by looking at the mean accuracy of each item for the two response groups in correspondence of the four classification methods (Figure 2), it is possible to observe that the number of items with accuracy lower than the chance level varies across the methods and in particular the normative and the fixed threshold methods show a quite high number of items with accuracy higher than 0.4.

Finally, to choose the RT threshold the RT distribution has been classified according to two characteristics: the presence of more than one mode detected by the Hartigan's Dip Test and the skewness of the curve. Figure 3 shows the mean accuracy grouped by achievement level for each item type in correspondence of each threshold method for the *rapid-guessing* responses. The mean accuracy of the *rapid-guessing* responses should be always close to the chance level and not associated with the achievement level. Overall, RA20 performs better when the distribution is not bimodal and not skewed whereas RA10 performs better when the distribution is left-skewed. One thing is worth to be highlighted is that in the ideal situation (bimodal and no skewed) the distribution of the accuracy by the achievement level is not affected by the choice of the threshold: for all the thresholds the accuracy of the *rapid-guessing* responses seems to be independent with respect to the competence level. The results reported in Figure 3 confirmed the need to choose a different threshold for each item based on the item characteristics.

Table 1. Accuracy of item responses for *solution behaviours* and *rapid-guessing* in correspondence of the four threshold methods.

Threshold	Solution Behaviour		Rapid-guessing	
	Accuracy	*N*	*Accuracy*	*N*
NT10	0.54	8163602	0.26	90559
Fix	0.54	8214217	0.23	36350
RA10	0.54	8193431	0.17	57136
RA20	0.55	8107583	0.19	142984

Classifying behaviours

Next, for each item, the EB index is computed. Among all the students' responses, 98.6% were identified as in *solution behaviours*. This percentage results to be higher than the percentage of *solution behaviours* reported for low-stakes assessments (94.2% in Wise and DeMars, 2006; 89.7% in Wise et al., 2009). Table 2 shows the mean RT and the mean accuracy for the two responses types. The mean accuracies of the two groups respect the two validation criteria: *solution behaviours* show higher accuracy than *non-effortful* and

the accuracy of *non-effortful* responses is close to the chance level. The behaviour most frequently identified was *rapid-guessing* (85.41%) whereas only 14.59% of the *non-effortful* response are classified as rapid omit.

Figure 2. Mean accuracy in correspondence of each item for *rapid-guessing* in red and *solution behaviour* in blue. The mean accuracy is computed for each item with the four threshold methods.

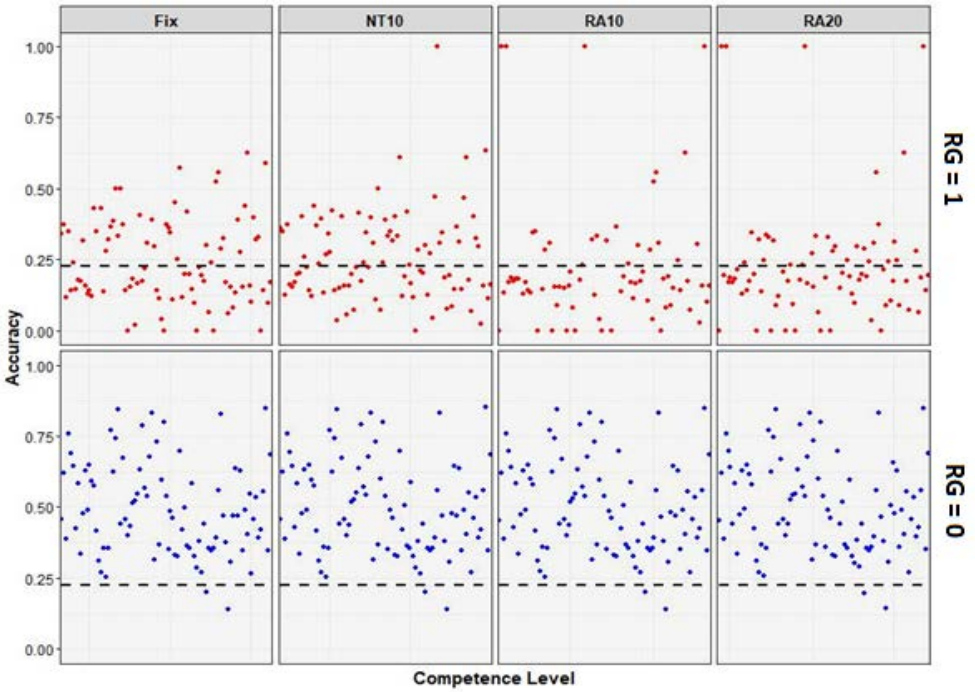


Figure 3. Mean accuracy of item responses for *rapid-guessing* response by the first three achievement levels.

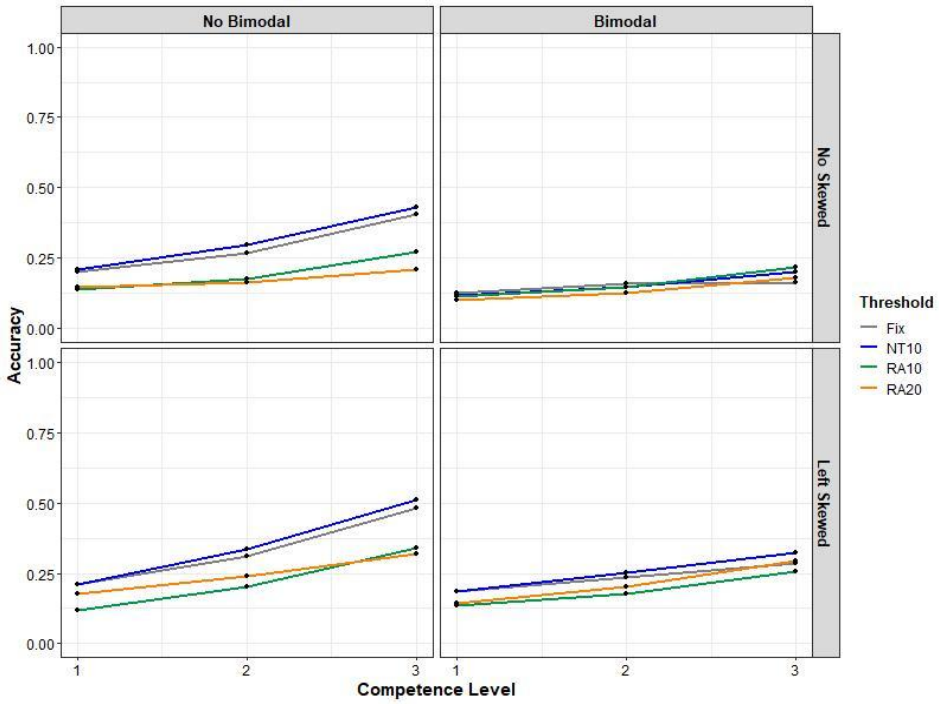
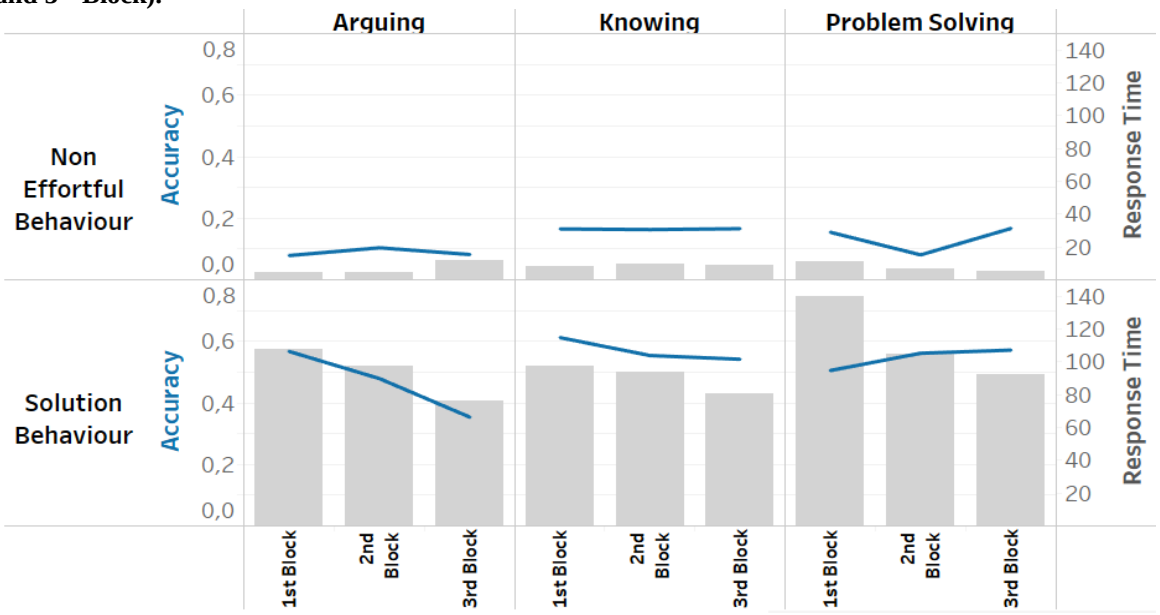


Table 2. RT and accuracy for *solution behaviours* and *non-effortful* responses.

Solution Behaviour			Non-Effortful Response				
Mean RT	***Accuracy***	***N***	***Mean RT***	***Accuracy***	***N***	***% RO***	***%RG***
95.53	0.56	1E+07	8.69	0.16	140475	14.59%	85.41%

The mean accuracy and mean RT of the two response groups are studied with respect to two item characteristics: the item dimension and the item position (Figure 4). In this analysis, the item position is categorized into 3 classes. To assess the reliability of the EB indicator, it was verified the last validation hypothesis: *non-effortful* responses should have an accuracy rate consistent with that expected from random responding regardless of the item position or dimension. The findings in Figure 4 support this hypothesis: for *non-effortful* responses the accuracy is always close to the chance level regardless of the item position or the item dimension. The accuracy of the *solution behaviours* responses is a decreasing function of the item position for the *arguing* and *knowing* dimensions. The mean RT in correspondence of the *solution behaviours* responses is always a decreasing function of the item position.

Figure 4. Mean accuracy and mean response time for *non-effortful* behaviour and *solution behaviour* with respect two different item characteristics: the item dimension (*Arguing*, *Knowing* and *Problem Solving*) and the item position (1st Block, 2nd Block and 3rd Block).



Response Behaviour Effort for students

RBE indicate the proportions of responses classified as effortful. The distribution of RBE scores, as shown in previous research, is negatively skewed, with RBE equal to 1 for most of the test-takers. In our data, 84.30% of the students have an RBE score equal to 1, whereas 3.35% of the students show an RBE score lower than 0.9. When the percentage of non-effortful responses exceeds 10%, the test performance tends to be materially

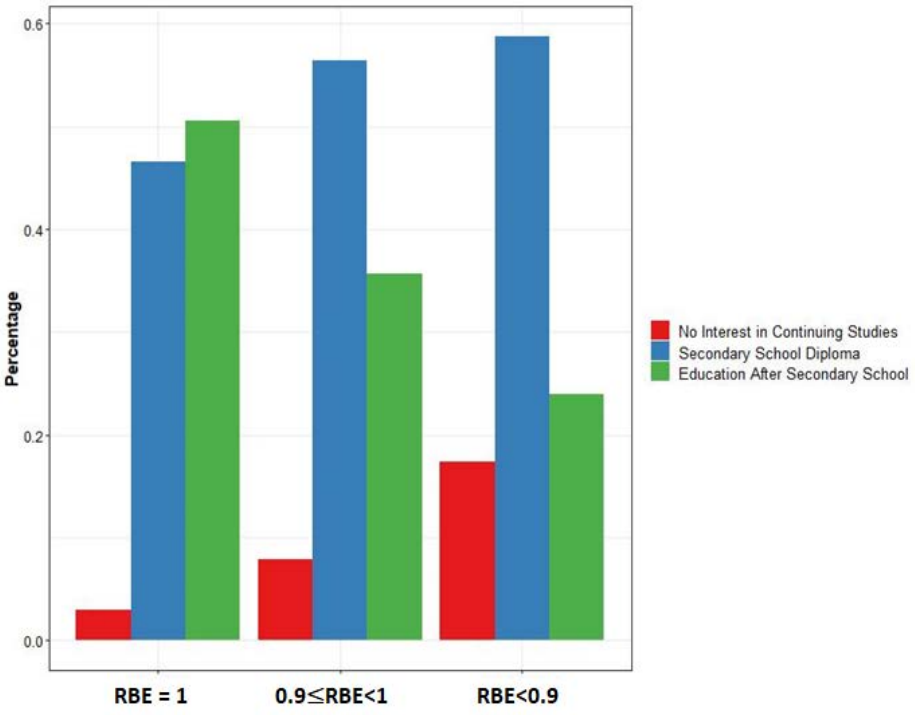
distorted (Wise, 2015; Wise and Kingsbury, 2016). Thus, in literature 0.90 is considered a useful criterion for invalidating a score because not trustworthy.

In addition, the distribution of RBE score is examined with respect to the response to a question of the Student Questionnaire administrated immediately after the math INVALSI test. The question analysed investigates the student educational expectation; in particular, the question was categorized in 3 classes: no interest in continuing the studies, interest in achieving a secondary school diploma and interest in continuing the studies after the secondary school. Figure 5 presents a clustered bar chart for the question analysed. The horizontal axis shows the three RBE categories and the vertical axis shows the percentage of each answer in each RBE category. It is evident that the percentage of the students answering that is not interested in continuing studies after the lower secondary school (in red) increases as RBE decreases, whereas the percentage of the students that want continue the education after the upper secondary school (in green) decreases as RBE decreases.

Table 3. Response Behaviour Effort (RBE) results for INVALSI data

	RBE=1	0.9≤RBE<1	RBE<0.9
All students	84.30%	12.34%	3.35%
Male	81.76%	13.89%	4.35%
Female	86.92%	10.75%	2.32%

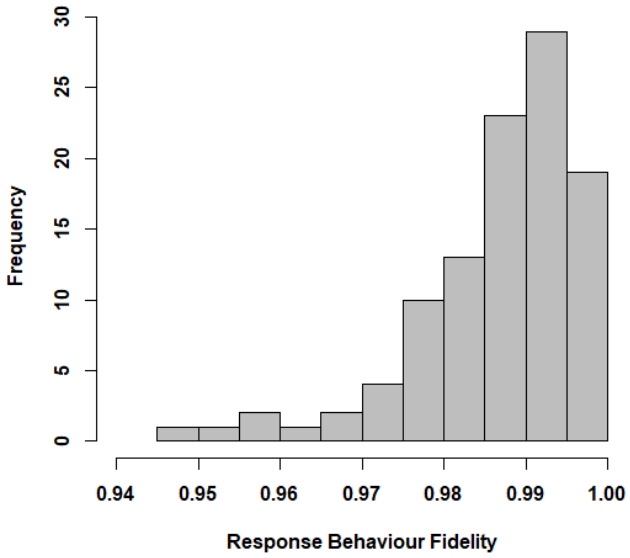
Figure 5. Clustered bar chart for the distribution of the questionnaire question about the student interest in obtaining a diploma conditional on the RBE levels.



Response Behaviour Fidelity (RBF) by items

The RBF represents the amount of effort received by each item. An item with higher RBF means that more students have shown an effortful behaviour on that item. Figure 6 shows the histogram of the effort received by the 145 closed-ended analysed items. The amount of the effortful responses ranges from 0.95 to 1, the maximum effort that an item can receive. Among the total of 145 items, only four items have an RBF smaller than 0.96, seventeen items have RBF ranged between 0.96 and 0.98, thirty-six item have an RBF lying between 0.98 and 0.99.

Figure 6. Histogram of the effort received by the 145 closed-ended items



Analysing the RBF of each item with respect to the position on which has been administrated (Figure 7), it is interesting to notice that there is a clear decreasing pattern: items appearing later tend to be characterized by smaller RBF. Six items in the third block had RBF smaller than 0.96, whereas in the second block and in the first block there are only one and zero items, respectively. All the items in the first block had an RBF higher than 0.99, whereas only six items in the third block. The decreasing relation between the effort received by a given item and the item position is confirmed by other previous works (Setzer et al., 2013; Lee and Jia, 2014). The effort received by the items has been analysed related to the items' dimension. The boxplot of RBF (Figure 8) suggests that the item of the *arguing* dimension had higher RBF than the items of the other two dimensions. The linear mixed model analysis investigates which item characteristic affects the RBF; three items characteristics were used as predictors: item difficulty, item position and item dimension. In this analysis, the item position is not used as a categorical variable but as a continuous variable. The p-values (Table 4) confirmed the findings of the descriptive analysis.

Figure 7. Response behaviour fidelity by item position (1st Block, 2nd Block, 3rd Block).

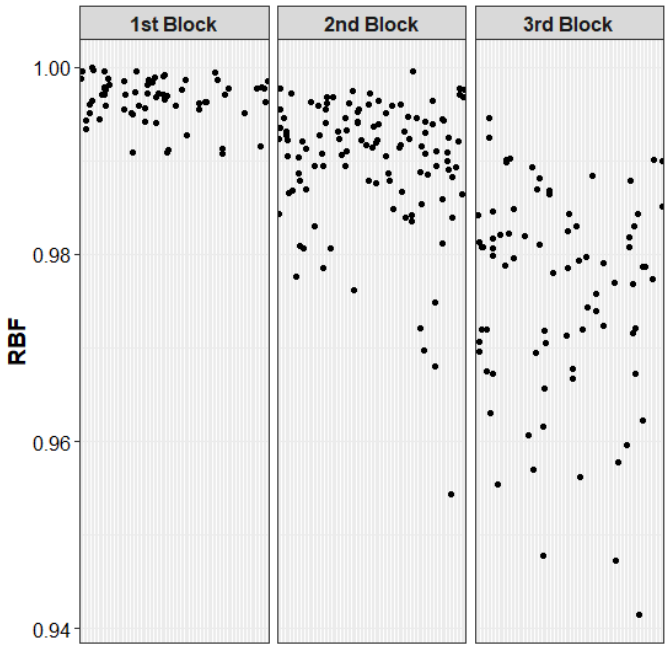


Figure 8. Boxplot of response behaviour fidelity by item dimension.

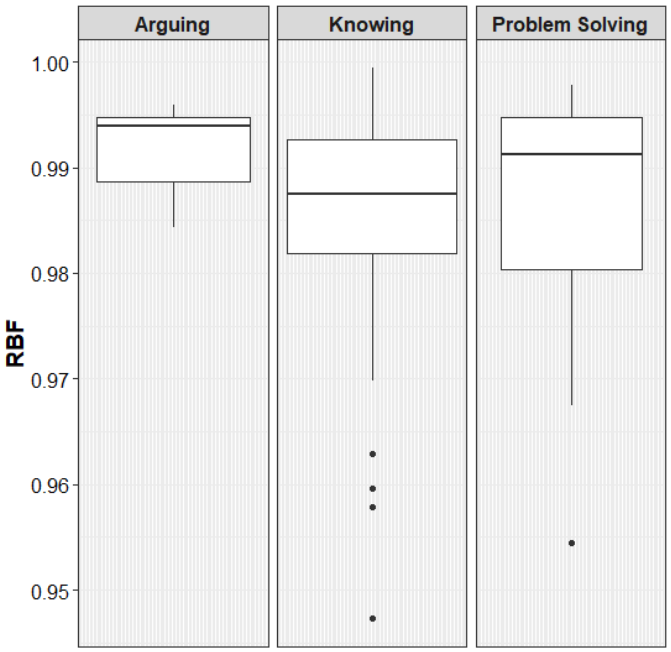


Table 4. Results of linear mixed model. Analysis of response behaviour fidelity

	Coefficient	Std.Error	p-value
Intercept	0.992	0.003	<2E-16
Position	-0.001	0.000	<2E-16
Item Difficulty	-0.008	0.003	0.003
Dimension: Problem Solving	0.014	0.006	0.015
Dimension: Arguing	0.012	0.008	0.142

Conclusions

In this work, a preliminary study of the response time of INVALSI data was performed. A new procedure to examine the test-taking behaviours using response time was proposed; in particular, it provides a way for the identification of the rapid-guessing and the rapid-omit. The computed effort indicator shows different advantages: the indicator is based on the response time that is collected unobtrusively for each student-by-item interaction; the indicator can be summarized at both item level and student level and this allow to study the engagement with respect to both item and student characteristics; the indicator provides a measure of the student behaviour over the entire test event.

This paper reveals, for the first time, a measure and the extent of the engagement during the INVALSI assessments of grade 8. A very low percentage of non-effortful behaviour was identified in our data, however Wise and DeMars (2006) demonstrated that modest amounts of *rapid-guessing* behaviour can affect the precision and the reliability of the scores. In future studies, a *motivation filtering* procedure, that consists of removing the non-effortful examinees, will be applied to analyze the impact on test scores of the *non-effortful* responses (Sundre and Wise, 2003). An agreement between the student's engagement measured using the response time and the students' educational expectation has been found and should be further investigated in the next studies. One of the most interesting findings is related to the analysis of the response effort at item-level. The effort received by items is negatively associated with the item position and with the item difficulty. These findings are consistent with the literature (Wise et al., 2010; Seitzer et al. 2013; Lee and Jia, 2014): the student's disengagement tends to emerge during the test event and consequently the quality of test data decreases.

The findings of this study have two principal limitations: the measure of the engagement is based on only the closed-ended response and the classification of item responses as non-effortful has done conservatively. Additional research should conduct to identify effective criteria for the identification of the non-effortful responses of the open-ended items and to choose the RT threshold for the rapid-guessing responses. Future research will conduct to shed light on these two points. Moreover, it is well known that the disengagement occurs most frequently with low-stakes tests, for this reason in future works the extent of the engagement to INVALSI assessment will estimate for grade 10.

References

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment. *Teachers College Record*, 113(11), 2309-2344.

Breakspear, S. (2012). The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. OECD Education Working Papers, No. 71. *OECD Publishing (NJ1)*.

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

Fullan, M. (2009). Large-scale reform comes of age. *Journal of educational change*, 10(2-3), 101-113.

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

INVALSI (2019). Rapporto prove invalsi 2019. https://invalsi-areaprove.cineca.it/docs/2019/Rapporto_prove_INVALSI_2019.pdf

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, 67(4), 606-619.

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, 2(1), 8.

Ma, L., Wise, S. L., Thum, Y. M., & Kingsbury, G. (2011, April). *Detecting response time threshold under the computer adaptive testing environment*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, 26(1), 34-49.

Sundre, D. L., & Wise, S. L. (2003, April). Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests. Paper presented at *Annual Meeting of the National Council on Measurement in Education*, Chicago, IL.

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, 2(3), 1–17.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, 28(3), 237-252.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, 36(4), 52-61.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment*, 10(1), 1-17.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30(4), 343-354.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, 53(1), 86-105.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego

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

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*

. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). An Investigation of the Relationship between Time of Testing and Test-Taking Effort. *Northwest Evaluation Association*.

Appendix

Figure A1. Histogram of response time and accuracy for an item with clear bimodal pattern (Hartigan's Dip Test p<2.2E-16, Skewness = -0.3). The item difficulty is low and the accuracy tends to increase as the RT increases. The x-axis is truncated at 100 seconds.

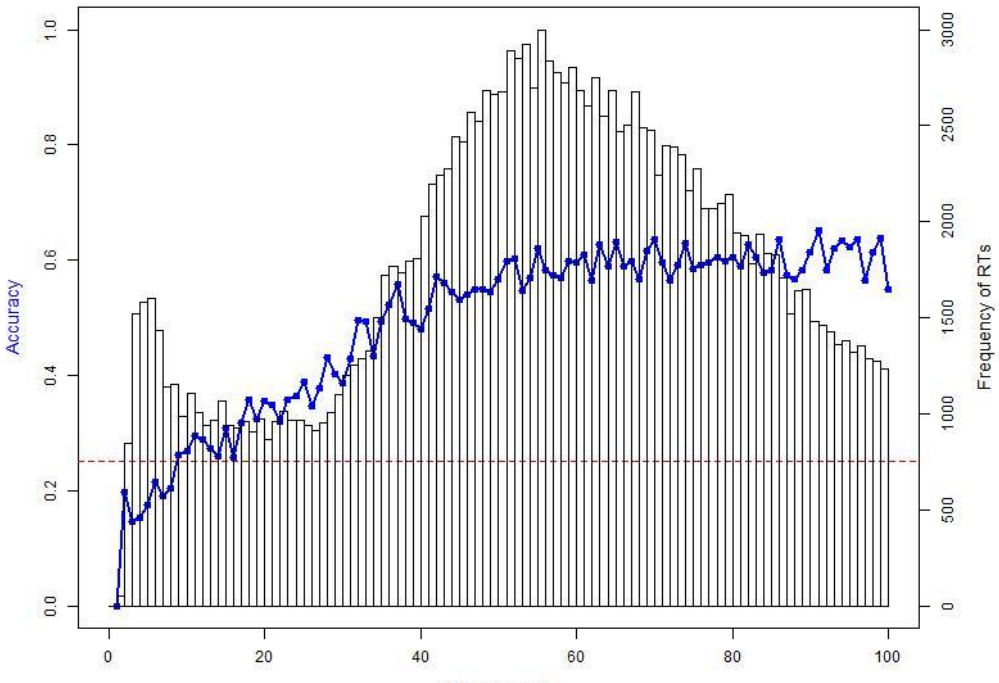


Figure A2. Histogram of response time and accuracy for an item with a heavy right tail (Hartigan's Dip Test p = 0, Skewness = -0.3). The item difficulty is high and the accuracy tends to decrease as the RT increases. The x-axis is truncated at 100 seconds.

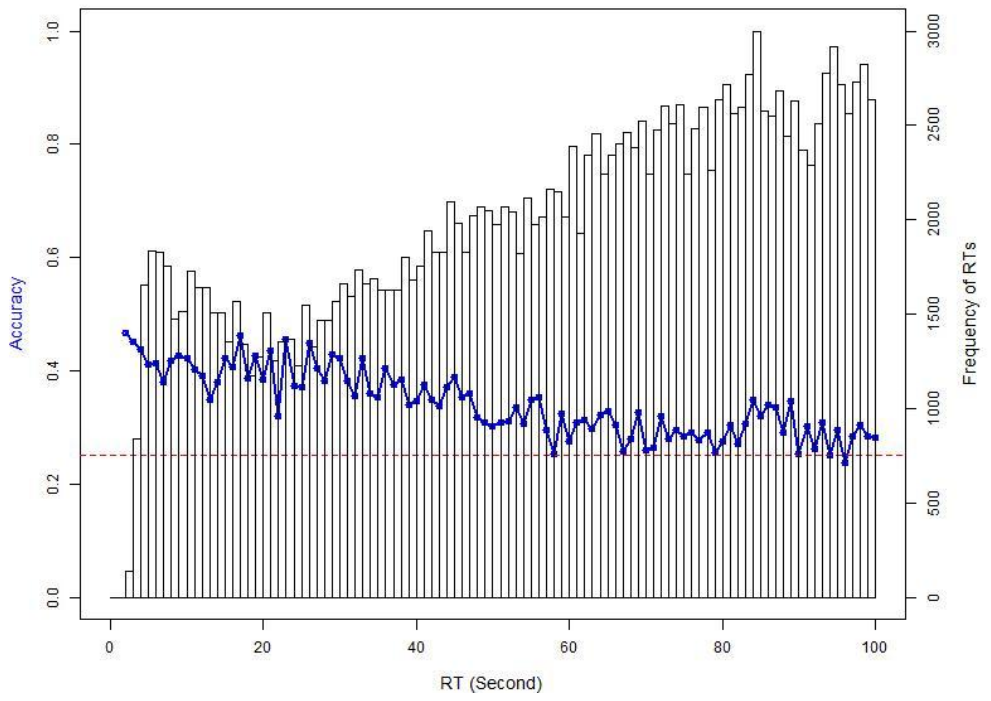


Figure 3. Histogram of response time and accuracy for an item with a long right tail (Hartigan's Dip Test p = 1, Skewness = -0.79). The item difficulty is high and the accuracy tends to increase as the RT increases. The x-axis is truncated at 100 seconds.