Speeding up without Loss of Accuracy: Item Position Effects on Performance in University Exams

Leonardo J. Vida
Utrecht University
Heidelberglaan 8
Utrecht 3584 CS
The Netherlands
l.j.vida@uu.nl

Maria Bolsinova
Tilburg University
Prof. Cobbenhagenlaan 225
Tilburg 5037 DB
The Netherlands
m.a.bolsinova@uvt.nl

Matthieu J.S. Brinkhuis
Utrecht University
Princetonplein 5
Utrecht 3584 CC
The Netherlands
m.j.s.brinkhuis@uu.nl

ABSTRACT

The quality of exams drives test-taking behavior of examinees and is a proxy for the quality of teaching. As most university exams have strict time limits, and speededness is an important measure of the cognitive state of examinees, this might be used to assess the connection between exams' quality and examinees' performance. The practice of randomization within university exams enables the analysis of item position effects within individual exams as a measure of speededness, and as such it enables the creation of a measure of the quality of an exam. In this research, we use generalized linear mixed models to evaluate item position effects on response accuracy and response time in a large dataset of randomized exams from Utrecht University. We find that there is an effect of item position on response time for most exams, but the same is not true for response accuracy, which might be a starting point for identifying factors that influence speededness and can affect the mental state of examinees.

Keywords

exam quality, computerized testing, item response time, item position effect, speededness, speed-accuracy trade-off

1. INTRODUCTION

The quality of standardized high-stakes tests can be seen as a driver of test-taking behaviors and mental states of test takers. The structured format of these tests, with strict time limits and high consequences attached to the test result, lead test takers to a situation in which they feel more or less comfortable [19]. With the introduction and spread of computerized testing in high stakes tests, more data can be collected on high-stakes tests than in the past. These data can be used to monitor the quality of measurement instruments of individual items and exams as a whole. Among the collected data, an important source of information are response times. Response times can be used to gain more information on the test-taking behavior of the examinees [17, 1] and the functioning of the exam and exam questions. As a proxy for exam quality, higher education institutions commonly use reliability measures such as the Cronbach's alpha [5], although literature showed that this indicator, if speededness is present in the exam, might be underestimated leading to reliability concerns [2]. However, reliability is not the only measure of exam quality: test takers behavior, in particular speededness, might be relevant to lecturers and test creators. Thus, investigating the presence of speededness in a test is not only important to know whether the commonly used reliability measure can be trusted, but it can also be used to propose a new indicator of exam quality that takes into consideration the cognitive state of examinees, relating tests' quality and examinees' performance.

Traditional measures of speededness only take into account whether examinees provide responses to all exam questions and are not missing a large proportion of items at the end of the exam [13]. However, fully missing responses at the end of the test is not the only way in which speededness manifests itself [16]. An important way in which time pressure can be observed is the increase of speed and decrease of accuracy close to the end of the test [11]. This behavior can be operationalized as the effect of item position on response time and response accuracy. When exam items are administered to all students in the same order, as often is in the case of traditional high-stakes achievement tests, the effect of item position cannot be separated from the effect of item properties. However, since for test security reasons exams are now more often administered with a randomized item order, it becomes possible to study item position effects separately from item effects.

We have a large data set of computerized exams administered at Utrecht University between January 2015 and June 2020. Using these data, we want to study the overall effect of item position on response time and response accuracy. Furthermore, for each exam we want to quantify the effect of item position on test performance which can be used as an indicator of test quality and of the mental state of test takers. To answer these questions, we focus on three key points. First, we uncover that responses to later items in exams have an increased speed, in conformity with previous studies on anxiety and test strategies within high-stakes

Leonardo Vida, Maria Bolsinova and Matthieu J. S. Brinkhuis "Speeding up without Loss of Accuracy: Item Position Effects on Performance in University Exams". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 454-460. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

tests [19, 8]. Second, we notice the lack of a relationship between item position and accuracy that, if analyzed in parallel with our first finding, might give us indirect evidence that increased response speed does not seem to have the expected negative effect on accuracy. This might show that successful test-taking strategies might include increased response speed towards the end of a test, as previous qualitative analyses already seem to show [18, 10], and that performance-reducing mental states do not appear to influence response speed.

1.1 Problem Statement

Currently, most tests used in European higher education institutions are non-adaptive computerized tests, that often have a large number of multiple-choice items, no penalty for incorrectly responded items and use a test-based time limit instead of a section-based time limit, as is usual in adaptive assessments. High-stakes non-adaptive computerized tests have not been researched and investigated as frequently as their adaptive counterparts and datasets on these tests are not widely available. Thanks to the advance of computerized testing within Dutch higher education institutions, we now have a multitude of data available that were not available before concerning high-stakes tests. Among these data, response times for each examinee on each item and the (randomized) item positions are saved. As test developers, when developing their tests, must find a balance between testing time requirements and difficulty and given that this balance depends on the type of the test and the needs of lecturers and students, we believe that exam-specific effects of item position on response time and response accuracy can provide them with useful information. Therefore, using the dataset at hand we set out to answer the following questions:

1. What is the effect of item position on response time and accuracy?
2. What can be inferred concerning test-taking behavior by analyzing the influence of item position on response time and accuracy?

1.2 Contribution

Answering our research questions, we make several contributions to the field: (1) using generalized liner mixed models that make use of item position in a dataset of university exams, we analyze the effect of item position on response time and accuracy; (2) we provide indications concerning the relation between examinees' response time and response accuracy within randomized high-stakes tests.

This paper is structured as follows: in section 2, we provide the background from which this work stems. Section 3, describes the data and the models used in the analysis. Section 4 continues comparing the results of the models fitted. Section 5 discusses the results of the models and provides the ground for the conclusion in section 6.

2. BACKGROUND

This research revolves around data collected at a higher educational institution concerning the results of tests administered using computers. The computerized collection of students' answers enables the creation of a dataset containing, among other data, the response time data of each student on each test item. We want to make use of this information to help us better understand response processes and, as a consequence, improve measurement instruments.

Interest in response time as a method of revealing information about mental activity has a long origin [15] and other research aims to be relatively comprehensive on the domain [6]. Here, we focus on the specific features on the approach relevant to our data and findings. Recently the role of response time modeling rose to a central position with novels works on the interplay between accuracy and response time [21, 9] and on item position [7]. Traditionally, two main effects of item position have been distinguished: a learning effect when items in later position become easier and a fatigue effect when items become instead more difficult [7]. In both cases, item position effects refer to the impact of the position of an item within an exam on the response time and on the response accuracy. Research commonly assumes that an increase in the speed of response will result in a decrease in accuracy [9]. This relationship, called speed-accuracy trade-off (SAT), is understood as a within-person phenomenon in which the accuracy of response varies with the time taken to produce it [11]. Our empirical findings provide some evidence that might enhance our understanding of the SAT and specify cases in which this relationship is more unclear than what previously thought.

On the other hand, the psychology and education literature has long been interested in developing test designs that generate fair results and thus studied examinees' test-taking behavior to investigate the effect of test designs. Among many domains, this literature also focuses on the effects of anxiety, motivation and test-taking strategies on performance when taking a test [19, 8, 3], finding that high achievers are more likely to engage in effective test-taking strategies compared to low achievers and identifying differences between genders in risk-taking behaviors and anxiety levels. Studies in this area identify risk-taking as an important strategy when taking a test, in particular in multiple-choice tests under strict limits [3]. These guessing strategies are found to potentially lead to better results regardless of ability level and compared to students at the same ability level not using these types of strategies [8]. Our empirical findings provide some evidence also in this aspect, not finding a negative relationship between speeded behavior and response accuracy.

3. METHODS

3.1 Data

We use data from Utrecht University that comprises all exams carried out using the online platform *Remindo Toets*[1] between 2015/01/01 and 2020/06/01. Given our goal of investigating the effect of item position, we select exams in which *response randomization* was applied (i.e., the position of the questions given to examinees changes from examinee to examinee). Therefore, the starting dataset of exams is filtered on the following conditions: (1) duration of the exam: less than 240 minutes. (2) Number of examinees: at least 100. (3) Number of items per examinee: at least 10. (4)

[1] *Remindo Toets* is a software product developed by Paragin, a Dutch education company, which provides educational institution with a platform to create, administer, review and grade exams.

Types of questions in the exam: "choice", "inline-choice", "order", "match". (5) Maximum response time: less than 600 seconds, to reduce outliers in the dataset. (6) Finally, we only analyze exams in which the item order is fully randomized. After filtering, the dataset contains 599.519 item responses. In the final dataset, tests are composed by an average of 204 students, 34 items and the average duration is of 106 minutes.

For each question, lecturers are provided with the so-called p-value, as a measure of "difficulty" of the item, and the estimates of commonly used metrics the item-test correlation (RIT) and the item-rest correlations (RIR) [20]. These variables are available along with item responses and response times. On the dataset at hand, the mean response time is of 91.18 seconds while the average accuracy rate is of 63.19%.

Due to privacy reasons, this dataset was anonymized. Because of this, we are only able to provide a general overview of the exams used in this analysis and not a complete overview of the underlying students population. The dataset at hand consists of 90 unique exams across 6 faculties within Utrecht University. In order, the largest faculties are Science, Veterinary Science and Social Sciences. Finally, as we selected courses with more than 100 examinees and due to the difference in the average class size between bachelor and master courses at Utrecht University, the wide majority of selected exam were from bachelor programs which are typically taught in Dutch. The predominance of the exam in Dutch (90%) indicates that the majority of students attending these courses courses are Dutch, implying that the results of these analyses might be culture-specific.

3.2 Selected variables

In the context of our analysis, we make use of the following variables:

- *Student*: factor variable identifying each individual student. Total number of factors: 18.476.
- *Test*: factor variable identifying each individual test. Total number of factors: 90.
- *Item*: factor variable identifying each individual item. Total number of factors: 5.089.
- *Response time*: continuous variable referring to the total time, in seconds, spent by an examinee on an item. The response time is the summed response time across all attempts made in answering that item. Clear extreme outliers in item were eliminated by setting a cutoff in the filtering process of the data.
- *Accuracy*: binary variable referring to a right or wrong answer by an examinee on an item.

Additionally, we create the following two variables: *item position* and *available time per item*. The first variable, *item position*, is used to identify the location in which the item appears within a test and it is divided within 10 blocks representing 10% of the exam. For each response of person i to item j in exam k, $z_{ijk} \in [0:9]$ denotes the block in which the item was presented to the person. We also create a set of dummy variables $z_{1ijk}, \ldots, z_{9ijk}$, where z_{sijk} if $z_{ijk} = s$, in order to model nonlinear relationship between item position and exam performance. The second variable, *available time per item*, is created dividing the total allotted time for a specific exam by the number of items in that exam. This variable is created as the exams available in the dataset are heterogeneous and do not have a common time limit. As we cannot compare exams having different time limits, we create a variable that represents the time limit at an item level. Across all exams, the *available time per item* has a mean allotted time of almost 3 minutes (177 seconds).

3.3 Models

Before discussing the results of our models, we make a brief note of the reason underlying their creation. A key necessity in our models is the ability to quantify the effect of item position on response time and on response accuracy. Therefore, to build models we turn to generalized linear mixed models (GLMMs). We choose GLMMs as we aim to develop models that create a reliable and easily repeatable analysis to increase the reach and applicability of the model results to datasets from other educational institutions.

In both models, to study the effects on response accuracy and response time, we consider three predictors. We allow for random variability across students incorporating this effect as random effect (θ_{1ik} for the effect on response accuracy, and θ_{2ik} for the effect on response time). We consider fixed item effects (β_{1jk} and β_{2jk} for the effects of item j from exam k on response accuracy and response time, respectively). Finally, for each of the item position dummy variables we consider fixed effects on response accuracy (γ_{1sk}) and on response time (γ_{2sk}), which are estimated for each exam separately. We include fixed item position effects only in the second variation of both models in order to enable us to evaluate whether their addition is significant.

3.4 Modeling response time

We construct the models concerning response time using the logarithm transformation of response time. For response time, we focus on the following linear mixed effect regression ($LMER$) models:

$$\begin{aligned} y_{ijk} &= \beta_{2jk} + \theta_{2ik} \\ \theta_{2ik} &\sim N(0, \sigma^2_{1k}) \end{aligned} \tag{1}$$

$$\begin{aligned} y_{ijk} &= \beta_{2jk} + \sum_{s=1}^{9} \gamma_{2sk} z_{sijk} + \theta_{2ik} \\ \theta_{2ik} &\sim N(0, \sigma^2_{2k}) \end{aligned} \tag{2}$$

wher y_{ijk} is the log-transformed response time of person i on item j in exam k, and σ_{2k} is the variance of the person random effect on response time in exam k. Model 1 does not contain the fixed effects of item position.

3.5 Modeling response accuracy

For response accuracy, we focus on the following generalized linear mixed effect regression on a binary variable (GLMER):

$$\begin{aligned} logit(x_{ijk}) &= \beta_{1jk} + \theta_{1ik} \\ \theta_{1ik} &\sim N(0, \sigma^2_{1k}) \end{aligned} \tag{3}$$

where x_{ijk} denotes response accuracy of person i on item j in exam k, and σ^2_{1k} is the variance of the random effect.

The model is extended with the effect of the item position dummy variables:

$$logit(x_{ijk}) = \beta_{1jk} + \sum_{s=1}^{9} \gamma_{1sk} z_{ijks} + \theta_{1ik} \qquad (4)$$
$$\theta_{1ik} \sim N(0, \sigma^2_{1k})$$

4. RESULTS

We first analyze the results of response time models from equation 1 and 2, before moving to the results of the response accuracy models from equation 3 and 4. Because of limitations due to the size of data at hand, in particular due to the number of fixed effects, all models are fit on each exam individually and their effects are shown as the gray lines in Figure 1 and Figure 2. After fitting each individual model, the item position effects estimates are pooled together with a random-effect meta-analysis in which we the exam-specific effects are assumed to come from a distribution with a common mean and variance [4]. The means of the effect and the ± 1.96 times the standard deviation of the effect boundary are shown as the blue and light blue lines in Figure 1 and Figure 2. The estimates of each of the item position effects of the pooled model of *LMER 2* and *GLMER 4* are given in Table 1 and Table 2 in the Appendix.

4.1 *LMER* on log response time

Concerning *LMER* models on response time, we see that across exams the ANOVAs between model 2 model 1 are significant in 79 out of the 90 cases. Respectively 72 exams at the .1% significance level, 5 at the 1% significance level and 2 at the 5% significance level. The average additional proportion of explained variance of model 2 on model 1 is 0.0084.

Analyzing response time behavior and the effect of item position, we observe a significant negative effect of item position on response time. In particular, we observe an increase in the effect over the course of the exam, which is typically defined as response acceleration. This can be seen in Figure 1 as each item position relates to the effect size compared to the baseline of item position 0, which represents the first 10% of the exam. The blue line represents the pooled estimates of the item position effects across all exams, while the gray lines represent individual exams.

4.2 *GLMER* on response accuracy

With respect to *GLMER* models on response accuracy, we see that model 4 does not significantly outperform model 4. The ANOVAs between model 4 and model 3 are only significant 5 times at the 1% significance level and 11 at the 5% significance level. This shows that the response acceleration found previously and shown in Figure 1 either is not strong enough to influence response accuracy or does not have any influence on it. Hypothesis on the reason underlying this finding are discussed in section 5.

4.3 Proportion of explained variance and available time per item

After running a likelihood-ratio test on each exam individually, we select model 2 as significantly better than model 1 while we further investigated the additional proportion of explained variance of model 2 regressing it on the *available time per item* for each exam. Figure 3 shows that there is a relationship between the available time per item in an exam and the additional explained variance from the model with item position effects (linear correlation $-.37$). This is an indicator that on exams that have less time available, response acceleration is indeed happening and the inclusion of a variable to take into account the position of the item help us explain better the behavior of students. However, as it is visible in Figure 3, the additional explained variance of model 2 is relatively low. This result is important as it provides us with a tool to support the results of response acceleration.

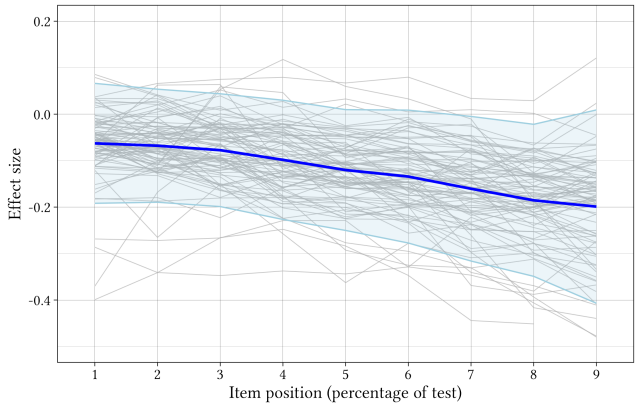

Figure 1: Item position effects on response time

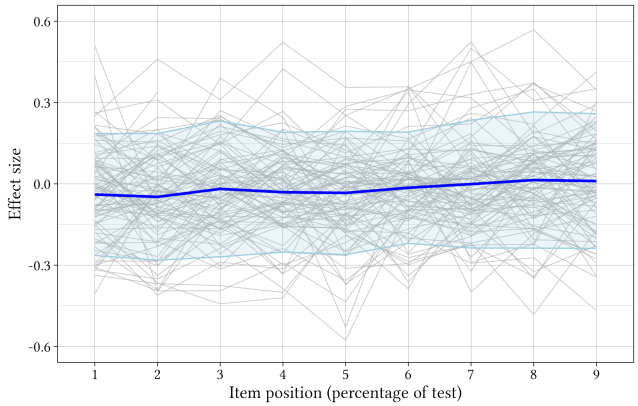

Figure 2: Item position effects on accuracy

5. DISCUSSION

5.1 Discussion of findings

Using a collection of item responses and response times from higher education tests during a five-years time span, we analyze the relationship between item position and response time and between item position and response accuracy. We show that item position is associated with response acceleration while we find that the connection between item position and response accuracy is unclear. Finally, we also find that the available time per item is negatively correlated with the additional explained variance when comparing the model relating item position and response time.

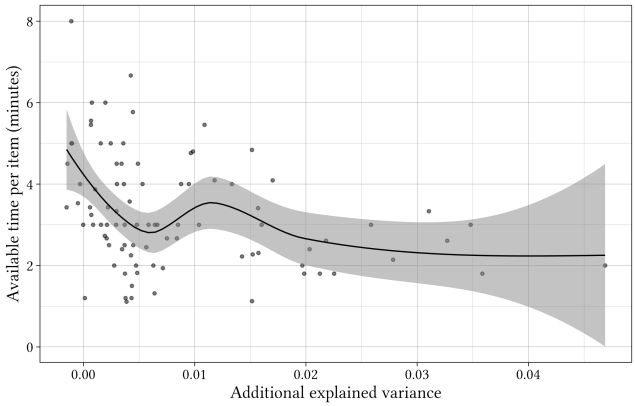

Figure 3: Regression between mean duration of item and additional explained variance for model 2

Concerning the significant effect between item position and response times, we see that the effect is not as strong as we initially expected. The presence of this effect might stem from increased respondents' fatigue or decreased interested in the exam, as highlighted by previous work on this topic [22, 9]. However, previous literature focuses on adaptive tests in which items becomes more or less hard in relation to respondent's performance, while in our dataset this is not the case. Comparing the differences between these two modalities of testing and understanding the difference in the response behavior might be an attractive opportunity for future research.

With regards to the interaction between response accuracy and item position, literature tends to show that later items are more difficult [14, 22] and therefore might decrease the rate of correct responses. We find no significant relationship between response accuracy and item position. The lack of this relationship might be explained by a few hypotheses. First, we might hypothesize that the response acceleration shown is the representation of students reaching their natural speed on the exam. Students might take some time to enter in the mental state that allows them to take an exam and they might be initially slowed down by the need of understanding how the exam is structured. This might explain the increased acceleration of response at the same level of accuracy between earlier and later item in the test. Secondly, we might hypothesize that there is an negative effect of response acceleration on response accuracy, but it is not shown because of the relatively heterogeneous dataset of exams at hand and because of the need of accessing more data. Concerning the second hypothesis, more work on a larger dataset of similar exams is needed before drawing any conclusion. Finally, the lack of this relationship might also be explained by the presence of increased response speed within effective test-taking strategies among high achievers, as found by previous qualitative literature on this topic [18, 12]. This result might be caused by an increased willingness to guess using effective elimination processes, leading to an effective guessing strategy on high-stakes multiple choice tests, when compared to the choice of picking an answer option at random.

Finally, we demonstrate that there is a relationship between the additional explained variance and the available time per item. When the model with item position effects is significantly more informative than the model without these effects, this correlation might also provide backing to a potential quality indicator to be provided to lecturers and test creators to inform them about their tests. In the presence of high additional explained variance, an indicator that would take this relationship into account might be used to provide lecturers with information about the quality of their tests and to identify exam-specific factors that influence speededness and the test-taking strategies of examinees.

5.2 Limitations

When carrying out the modeling part of our research, due to the size of data and factors at hand, we realized that the current statistical methods available to analyze this quantity of data create computational problems. As a matter of fact, we were interested in analyzing the effect of item position on response time and accuracy across the entire dataset but, due to computational limitations, we decided to fit the models on individual exams and later pool the effects estimates using a meta-analysis. To avoid this obstacle, two parallel path might be taken: (1) extending the current statistical libraries to include the possibility of using sparse matrices in computing fixed effects estimates and (2) adding more computational power to the tools used in the analysis.

Further, we also need to take into consideration both the dataset used in this research and the filtering actions taken on it. First, the dataset stems from a higher education institution (*wetenschappelijk onderwijs*) and therefore the results of our analyses might be dependent on the educational level of the students' population. Secondly, because of the filtering actions carried out on the dataset (2), we can assume that Dutch students are more represented in the dataset than international students. This might imply that the results stemming from our analyses are highly dependent on the Dutch test-taking "culture".

An important distinction between our work and previous studies on speededness, such as [9], is that we attempted to remove the effects of very long answers, but not of answers given during *rapid guessing behavior* [15]. As a matter of fact, the validity of test scores of such tests are threatened by what [23] call *noneffort*, which is associated with the guessing behavior of an examinee who does not try to solve items. The effect of such mechanism is an underestimation of his or her actual level of proficiency, threatening the validity of test score by adding a source of construct-irrelevant variance [23]. An analysis about the presence and the methods to identify *noneffort* in this dataset might reveal pathways to either take it into account or eliminate it from the dataset, paving the way for an analysis that compares the estimates of ability accounting only for *truthful* response acceleration.

5.3 Future research

Expanding dataset. As noted in section 5.2, we believe expanding the dataset at hand, including most recent data, and including other universities, would help providing more information on the relationship between response time and accuracy and thus the creation of more accurate indicators for lecturers. Moreover, as there might be differences in

the "culture" of examination between countries and between educational levels, a larger dataset comprising of multiple countries and different educational levels might help clarifying these questions.

Creating a metric to provide recommendations. As we did not see a clear effect of item position on accuracy, a future path for research might assume that, in the presence of similar results, students are given too few items for the time allotted on a specific exam. This would open a path for creating experiments to increase the total number of items on an exam and analyze the cutoff value at which effects on accuracy start appearing. An experiment increasing the number of test items would not only clarify the values at which effects appear, but would also improve domain coverage for that specific exam and, improving its reliability metric, and would allow us to study changes in test-taking strategies within the test.

6. CONCLUSION

This research evaluates methods to investigate the effects of item position on response time and accuracy. We find that, thanks to the advancement in the technologies used in exam settings and the wide application of these technologies in high-stakes tests, these analyses are not only feasible but are also promising if applied on larger datasets. We believe the overall results of the models can be of use within the educational sector, in particular thanks to the creation of an additional and reliable indicator of the quality of an exam. Using the results presented in section 4, we can now answer our research questions:

1. We find a small but significant effect of item position on response time, while we fail to find any effect of item position on accuracy. Section 5 provides a discussion on the findings.

2. We believe our results stemming from modeling item position and response time and the subsequent regression of the additional explained variance of this model and the available time per item, could evolve into a practical indicator providing more information concerning the quality of a test and the test-taking behavior of students.

7. REFERENCES

[1] A. P. Association. Standards for educational and psychological testing, 2014.

[2] Y. Attali. Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, 29(5):357–368, 2005.

[3] K. Baldiga. Gender differences in willingness to guess. *Management Science*, 60(2):434–448, 2014.

[4] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. *Introduction to Meta-Analysis*, volume 8. John Wiley & Sons, Bridgewater, NJ, USA, jan 2009.

[5] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

[6] P. De Boeck and M. Jeon. An overview of models for response times and processes in cognitive tests. *Frontiers in psychology*, 10:102, 2019.

[7] D. Debeer and R. Janssen. Modeling item-position effects within an irt framework. *Journal of Educational Measurement*, 50(2):164–185, 2013.

[8] H. Dodeen. Assessing test-taking strategies of university students: developing a scale and estimating its psychometric indices. *Assessment & Evaluation in Higher Education*, 33(4):409–419, 2008.

[9] B. Domingue, K. Kanopka, B. Stenhaug, J. Soland, M. Kuhfeld, S. Wise, and C. Piech. Interplay between speed and accuracy: Novel empirical insights based on 1/4 billion item responses, Mar. 2020.

[10] A. P. Ellis and A. M. Ryan. Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-eff icacy. *Journal of Applied Social Psychology*, 33(12):2607–2629, 2003.

[11] R. P. Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8:150, 2014.

[12] E. Hong, M. Sas, and J. C. Sas. Test-taking strategies of high and low mathematics achievers. *The Journal of Educational Research*, 99(3):144–155, 2006.

[13] Y. Lu and S. G. Sireci. Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(4):29–37, nov 2007.

[14] G. Nagy, B. Nagengast, M. Becker, N. Rose, and A. Frey. Item position effects in a reading comprehension test: an irt study of individual differences and individual correlates. *Psychological Test and Assessment Modeling*, 60(2):165–187, 2018.

[15] T. C. Oshima. The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31(3):200–219, 1994.

[16] D. L. Schnipke and D. J. Scrams. Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34(3):213–232, 1997.

[17] D. L. Schnipke and D. J. Scrams. Exploring issues of examinee behavior: Insights gained from response-time analyses. *Computer-based testing: Building the foundation for future assessments*, 34:237–266, 2002.

[18] T. Stenlund, H. Eklöf, and P.-E. Lyrén. Group differences in test-taking behaviour: An example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice*, 24(1):4–20, 2017.

[19] T. Stenlund, P.-E. Lyrén, and H. Eklöf. The successful test taker: exploring test-taking behavior profiles through cluster analysis. *European Journal of Psychology of Education*, 33(2):403–417, 2018.

[20] U. University. Toetsanalyse in remindo v2.1. `https://remindo-support.sites.uu.nl/wp-content/uploads/sites/79/2019/11/Toetsanalyse-in-REMINDO-v2.1-003.pdf`, Nov. 2019. Accessed: 2020-10-30.

[21] P. W. van Rijn and U. S. Ali. A generalized speed–accuracy response model for dichotomous items. *psychometrika*, 83(1):109–131, 2018.

[22] S. Weirich, M. Hecht, C. Penk, A. Roppelt, and K. Böhme. Item position effects are moderated by changes in test-taking effort. *Applied psychological measurement*, 41(2):115–129, 2017.

[23] S. Wise and C. DeMars. Examinee noneffort and the validity of program assessment results. *Educational Assessment*, 15(1):27–41, 2010.

APPENDIX
A. ESTIMATES OF POOLED MODELS

Table 1: Coefficients estimates of pooled model *LMER 2*

Item Position	Estimate	Std. Error	Estimate Rand.	Std. Error Rand.	tau
1	-0.0539	0.0036	-0.0628	-0.0628	0.0658
2	-0.0625	0.0035	-0.0679	-0.0679	0.0621
3	-0.0775	0.0036	-0.0775	-0.0775	0.0619
4	-0.0949	0.0035	-0.0983	-0.0983	0.0655
5	-0.1169	0.0035	-0.1203	-0.1203	0.0663
6	-0.1309	0.0035	-0.1343	-0.1343	0.0729
7	-0.1547	0.0036	-0.1605	-0.1605	0.0794
8	-0.1750	0.0035	-0.1854	-0.1854	0.0835
9	-0.1879	0.0034	-0.1989	-0.1989	0.1062

Table 2: Coefficients estimates of pooled model *GLMER 4*

Item Position	Estimate	Std. Error	Estimate Rand.	Std. Error Rand.	tau
1	-0.0412	0.0156	-0.0393	-0.0393	0.1148
2	-0.0412	0.0151	-0.0480	-0.0480	0.1193
3	-0.0163	0.0155	-0.0185	-0.0185	0.1279
4	-0.0283	0.0152	-0.0308	-0.0308	0.1127
5	-0.0264	0.0154	-0.0335	-0.0335	0.1161
6	-0.0106	0.0152	-0.0144	-0.0144	0.1046
7	0.0007	0.0154	-0.0009	-0.0009	0.1203
8	0.0141	0.0152	0.0141	0.0141	0.1282
9	0.0157	0.0147	0.0103	0.0103	0.1268