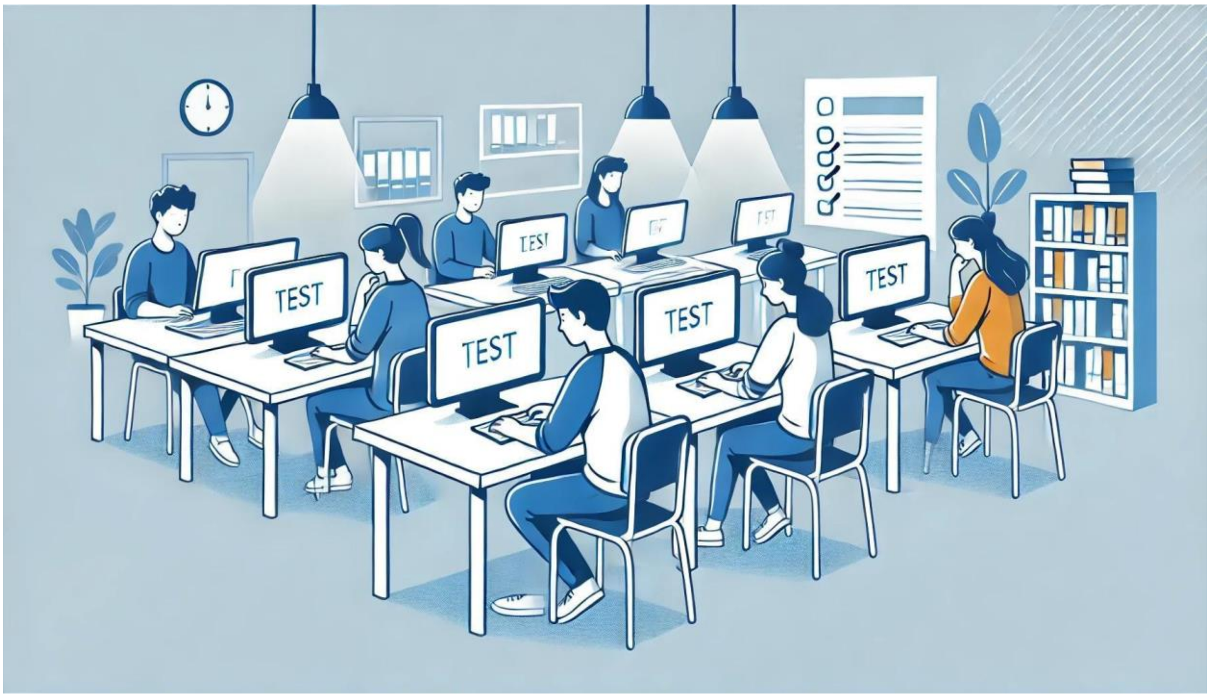
Technical report

IEA R&D Call 2023 grant:

Approximate Areas of Interest for Enhanced Understanding of Student Motivation and Task Interaction in IEA Assessments



Artur Pokropek, Tomasz Żółtak, Marek Muszyński

Institute of Philosophy and Sociology of the Polish Academy of Sciences

Contents

1. Background and aim of the study

The advent of computer-based assessments allows the collection of log data, a type of paradata that can be used to gain insights into testees' response behaviors and motivation. The complexity of traditional paradata, however, often limits its usefulness. Our team has developed a simplified, adaptable method, Approximate Areas of Interest (AAOIs), to capture specific screen areas where test-takers interact with the measurement tool (e.g. a test or a questionnaire). The approach is based on tracking cursor movements. It uses a flexible grid to identify distinct areas where respondents are most likely to interact with task elements on a screen. Using a vertical and horizontal splits system, the screen is divided into various Areas of Interest, such as item stem (introduction, information, instructions, request), response options (key and distractors, response area or scale with labels), and navigational buttons. In surveys, this approach has proven 95% accurate in identifying disengaged respondents in our previous research (Pokropek et al., 2024). In this project, we aim to adapt the method to the context of cognitive tests by determining AAOIs and providing an additional insight into item-solving behaviors.

This technical report offers a comprehensive summary of the project's implementation. It covers the design and development of the assessment design, assessment framework, and instruments (the cognitive test and questionnaire). The Technical Report details the outcomes of expert consultations undertaken during the project's development to achieve the highest standards of methodological design and provides information from qualitative and quantitative pilot studies conducted to validate the effectiveness of the research design in instrument design.

It also presents detailed results of the fieldwork study conducted in order to experimentally test the utility of mouse clickstream data under the framework of AAOIs to capture lower test motivation and related test behaviors (e.g. careless/insufficient effort responding).

2. Assessment Design

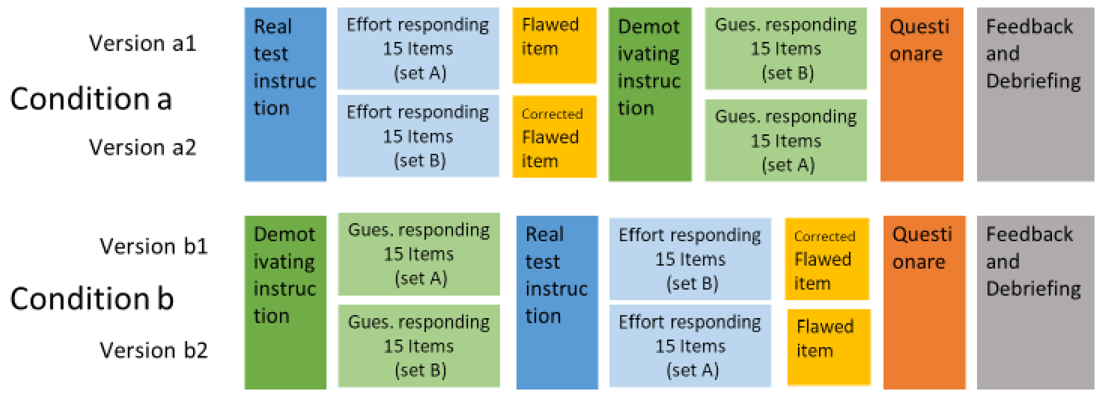
The study utilized tools from the TIMSS assessment for Grade 8, as well as Polish examination tasks designed for Grade 8 students for the primary school final exam. The target population consisted of students from Grades 7 and 8 within the Polish educational system. We ensured that all tasks were aligned with the curriculum for these grades. Including two cohorts in the study offered a significant advantage, as it facilitated the implementation of the research by allowing the study to be conducted in multiple classes within the same school.

The assessment planned to manipulate the participants' motivation to allow comparisons of response processes as captured by paradata patterns under high vs.

low test motivation. The manipulation entailed presenting different instructions in given test parts. In contrast to many similar studies (e.g., Huang et al., 2012; Pokropek et al., 2023; Rios, 2021; Ward & Meade, 2018; Ward & Pond III, 2015), this experiment used a within-participants design. In other words, each participant took a test under both conditions. Figure 1 presents the key elements of the assessment experimental manipulation plan.

Figure 1

Assessment experimental manipulation plan



The design incorporates two distinct conditions (Condition a and Condition b) to evaluate the impact of motivational manipulation on participant behavior and response patterns. The conditions are structured to compare the effects of motivational state changes induced by the instructions.

Condition a

1. **Real Test Instruction**: Participants receive initial instructions explaining the test and its objectives, setting the context for attentive responding. Instruction is available in Annex 1 (both in the original Polish and translated to English).

2. **Effort Responding Condition (15 items)**: This section requires attentive responding and assesses participants' response processes through 15 items. There are two versions of this test's part in which the same items appear, but in a counterbalanced order.

3. **Flawed Item**: A single item that is intentionally flawed to study participant reactions to apparent test errors or inconsistencies. The item appears in two versions: (a) flawed and (b) fixed. These items (translated into English) are presented in Annex 1 as well.

4. **Demotivating Instruction**: Following the real test effort condition, participants receive instructions intended to elicit careless/insufficient effort responding, measuring its impact on subsequent responses.

5. **Guessing Condition (15 items)**: In this section, participants are encouraged to guess the answers, allowing analysis of response processes under a simulated low-motivation condition. There are two versions of this test's part in which the same items appear but in a counterbalanced order.

6. **Questionnaire**: This final section gathers additional data about the participant's experience and perceptions throughout the test, as well as collects basic socio-demographic data.

7. **Feedback and Debriefing**: Provides participants with feedback on their performance and a debriefing with a rationale of the study.

Condition b

1. **Demotivating Instruction**: The test begins with instructions requiring inattentive responding, aimed at simulating response processes under reduced participant motivation. This is in contrast to the attentive responding instruction commencing the test in Condition A.

2. **Guessing Condition (15 items)**: As in Version A, this early segment gauges response processes under a simulated low motivation condition. There are two versions of this test's part in which the same items appear, but in a counterbalanced order.

3. **Real Test Instruction**: Midway through, participants are presented real test instructions to engage participants and "switch them" into an attentive responding mode.

4. **Effort Responding Condition (15 items)**: Similar to Version A, this segment gauges response processes under a simulated regular test motivation. There are two versions of this test's part in which the same items appear, but in a counterbalanced order.

5. **Flawed Item**: Flawed item examines if the sequence of motivational manipulation affects the reaction to test errors. The item appears in two versions: (a) flawed, and (b) fixed. These items (translated into English) are presented in Annex 1 as well.

6. **Questionnaire**: This final section gathers additional data about the participant's experience and perceptions throughout the test, as well as collects basic socio-demographic data.

7. **Feedback and Debriefing**: Provide participants with feedback on their performance and a debriefing with a rationale for the study.

It's important to note that the experimental design also incorporated a counterbalancing strategy regarding the item sets used in the test. This strategy ensured that each condition had two versions in which the placement of the "Effort Responding" and "Guessing responding" items was swapped between versions. This methodological approach was taken to control for any order effects or item biases associated with specific item sets. Therefore, the study follows the 2 x 2 within-objects design, with two versions of motivation condition (attentive vs. inattentive) and two versions of item order.

Manipulation check

The assessment also consisted of a manipulation check presented directly after each screen with instructions. The manipulation check was supposed to measure participants' attentiveness to instructions and gauge their understanding of the assigned task. The manipulation check looked as follows (translated version):

Before you start, we want to make sure you understand the instructions. How are you going to take on these tasks?

a) I have to answer the questions without putting too much effort into it, I can guess.

b) I have to solve the tasks as best as I can.

c) I should read them, but I'm not sure what to do.

d) I must ignore the instructions and act at my discretion.

Participants responded to this question to confirm their understanding of the experimental condition. If a participant's response did not correctly identify the experimental condition associated with the task descriptions, the instructions were presented again, and their understanding was re-assessed. This procedure was repeated until the participant provided a correct response. This iterative process ensured that each participant accurately understood the requirements of the tasks, thus maintaining the integrity of the experimental manipulations and the validity of the data collected. Such a procedure was used also to prevent skipping instructions.

3. Assessment Framework and Test Blueprint

The assessment framework was designed to align with the TIMSS Grade 8 math test and includes the following content domains: Number, Data and Chance, Algebra, and Geometry. The entire test is structured to reflect the Assessment Framework of TIMSS

Grade 8 and is supplemented by several tasks from the Polish high-stakes exams for Grade 8 (in Polish: *egzamin ósmoklasisty*). Scores from this exam play a prominent role in high school admissions. These items were joined with the TIMSS items to increase test motivation by presenting items similar to the ones used in the Polish Grade 8 final exam, which the target population will soon attempt. It was also believed that presenting the test as practice before the high-stakes exam would facilitate school participation. The test was planned to comprise approximately 30 items in total. Table 1 displays the blueprint of the test.

Table 1

Test Blueprint

Content domain	TIMSS Grade 8	Polish Grade 8 exam
Number	6 items	3 items
Data and Chance	4 items	2 items
Algebra	6 items	3 items
Geometry	4 items	2 items
Total	20 items	10 items

4. Instruments development

4.1 Math test

The original TIMSS Grade 8 items were selected with regard to the following criteria:

- Content that would allow to conduct all calculations in the students' heads as no calculators or paper notes were allowed during the test.
- Be of diverse difficulty so as to allow for a (close to) normal distribution of test scores. Balanced item difficulty was also important concerning the testee's motivation.
- Only multiple-choice items were selected to cut on time and costs of data collection and processing.
- Allow for optimal display on LimeSurvey online platform - for example none of the items employing graphical response could have been used.

The Polish Grade 8 exam items had to fulfill all of the above criteria, as well as:

- Represent the content domains of the TIMSS Grade 8 study.

The resulting test demonstrated good psychometric properties. Detailed item allocation and specific parameters of the final test can be found in Annex 2. However, it is important to note that the test used in this experiment is limited in both scope and format and does not fully replicate the actual testing conditions of neither TIMSS nor the national exam. This limitation should be taken into account when drawing inferences and interpreting the results.

The TIMSS items were translated into Polish by the project's PI whenever an official Polish version for an item was not available. Translated items were back translated by an English language translator with experience in educational assessment. The translations were then revised by the project's team with regard to the correctness of translation and natural language of the item.

All items were then implemented on LimeSurvey platform and tested internally to guarantee correct display of all of the selected items.

4.2 Questionnaire

The questionnaire was designed to minimize the time required for completion and to gather only the most essential information, reducing potential fatigue effects and maximizing the chances to fit in the testing session of 45 minutes - duration of the most typical school lesson in Poland. This time included all elements of the study - the test and the background questionnaire.

The questionnaire presented at the end of the assessment consisted of the following questions (response options in brackets):

- Grade (7 vs. 8),
- Gender (Male, Female, Do not want to answer),
- Self-reported test motivation,
- Strategy of responding in the "guessing condition" part of the assessment,
- The strategy of mouse use,
- Use of paper notes (Yes/No),
- Use of calculator (Yes/No),
- Use of telephone (Yes/No),
- Use of internet browser (Yes/No),

- Encounter of technical problems/glitches,
- Rating of how well the interviewer presented the assessment instructions,
- Number of books at home (a popular measure of socio-economic status, adapted from TIMSS Grade 4 background student questionnaire).

The questionnaire, like the test, was implemented in an electronic format using LimeSurvey assessment platform.

4.3 Assessment Procedures

Along with the tools, we developed specific assessment procedures. In our study, students completed the tasks in school computer rooms, in groups of around 20 and under the supervision of a teacher and an interviewer. School computer equipment was used for the assessment. The computer rooms were the same rooms regularly used for IT lessons, ensuring a familiar and comfortable environment for the students during the assessment.

Interviewers were responsible for ensuring that the appropriate equipment was available in the room as well as for presenting general test instructions for the students. Interviewers also guided students and provided technical support in case of problems.

A written form of the test standards and instructions was prepared, and each interviewer underwent a training session to ensure they were well-prepared to administer the assessment according to the standards and to handle any situations that might arise during the testing process.

5. Cognitive labs study

The fieldwork was preceded by a short cognitive interview study to assess whether instructions and procedures are understandable by the target audience. Eight students from grade 7 participated. A read-aloud protocol was employed — participants were asked to read the instructions aloud and then comment on each of the actions performed in the test. The project's PI, who performed all the sessions, recorded participants' actions and comments. Participants were tested individually.

No problems with understanding test instructions were noted. All items were understood well, and the level of correct answers varied between 40-75%, pointing to an adequate level of test difficulty.

Moreover, no technical problems were encountered with collecting the data – including log data – by the computer delivery platform.

Three mouse-using patterns were observed during the study:

a) Students use the mouse only to select the answer, without using it while reading.

b) Students use the mouse cursor to navigate the text, focusing on moving it while reading.

c) Students make mouse movements that are not directly related to the task (e.g., drawing circles), and only select the answer after finishing reading.

This fact inspired adding additional items to the main study's questionnaire for the main study, asking participants about the different strategies used to move the mouse while solving the test items.

6. Fieldwork procedures

The fieldwork procedures were designed to standardize the research, facilitate task execution for the students, and provide clear instructions to interviewers to ensure they can assist participants without disrupting the study. We have adapted these procedures from IEA studies and our previous assessments conducted in Polish schools. Special attention was given to the correct implementation of experimental stimuli. The same set of procedures was applied to both the pilot study and the main study.

6.1 Recruitment and Rewarding of Participants

The recruitment of schools was carried out by a recruited contractor - a specialized research company.

Entire classes participating in the study received small gifts from the research company in the form of minor educational aids, e.g. small electronic devices or office supplies. Individual students were not rewarded for participation. Instead, each student received information about their final test score and performance on each of the assessment's tasks. This information was displayed directly after the assessment.

Students in the final grades of primary school were preparing for their final exams after the 8th grade. The feedback on test results could be seen as a form of "reward," and interviewers emphasized the importance of this exercise, suggesting that students could use it as part of their Grade 8 final exam preparation. The final exam at the end of primary school is a high-stakes test that significantly influences the students' future opportunities.

6.2 Before the survey session

The interviewers were to enter the room at least 10 minutes before the session started and prepare it for ventilation, sunlight, an adequate number of seats, and appropriate spacing between workstations.

Minimum computer requirements were set to:

- A minimum 15.6” (inch) screen diagonal,
- A minimum 1024x768 resolution,
- RAM: at least 1 GB
- A processor with a frequency of at least 2 GHz, achieving a minimum score of 2500 points in the PassMark test,
- A full-size, optical, wired or wireless computer mouse included with each computer,
- An updated web browser that supports HTML 5.0 and JavaScript.
- Ensuring internet access with a minimum upload speed of 20 Mb/s throughout the interview.

If portable computers (laptops) were used for the study, a power supply and a battery that allowed for 2 hours of uninterrupted work without external power supply and a mouse must have been provided for each of them.

The interviewers were responsible to check the minimum requirements of the school computers and the presence and quality of the computer mice.

Screen resolution, as well as type and version of Internet browser was also read off by computer paradata collected by the testing platform. As the testing platform relied on JavaScript it was impossible to launch the test on obsolete computers.

On each computer designated for the survey, the interviewers navigated to the website *ankiety.ifispan.pl* and selected the link to the test version assigned to the current survey session. Then, interviewers were to enter the individual access code (token), which differed for each computer. The list of tokens assigned to a given session was prepared by the project’s team before the start of the fieldwork.

The interviewers also checked whether a mouse was connected to each computer and whether the device was turned on and fully functional. If there were laptops in the room, the interviewers ensured that students used mice instead of touchpads.

After logging in on each computer, the survey system displayed a welcome screen. The interviewers were to switch to full-screen mode by pressing the F11 key and invite the students to the room afterward.

6.3 During the survey session

After all the students had taken their seats at the computers, the interviewers ensured that there was exactly one person at each workstation and that no unauthorized persons were present in the room. Only students, teachers, and the interviewer were allowed.

The interviewers were to read the instructions to the students, emphasizing that:

- All mobile phones need to be silenced.
- No aids such as a phone, calculator, or paper notes are allowed.
- Students should work individually.
- Test instructions should be read carefully and obeyed.
- The tasks may appear in different order for different testees.
- The test is in full-screen mode and the students are not allowed to exit it.
- During the test students are not allowed to visit other websites.
- The *Submit* button needs to be clicked at the end of the survey.

The instruction also presented the test's structure (two parts with math tasks and a questionnaire at the end) and informed that after completing the survey, the testees would be able to view their results. The instructions also informed about the anonymity of the survey and that the results would be used solely for scientific purposes by researchers from the Polish Academy of Sciences.

The interviewers were also instructed to actively prevent any attempts to copy or record test items or to record the test session.

7. Fieldwork implementation

According to Polish law and IFiS PAN internal regulations, a tender was organized to select a company to conduct the fieldwork operations. The EDBAD Maciej Mroczek company was chosen among the total of six applicants on the basis of the price and other conditions offered.

Before data collection started, all participating interviewers took part in the 2h training session, during which participants were familiarized with data collection procedures, requirements, and technical details. The session included a presentation on the said issues and a Q&A session.

8. Pilot Study

A pilot administration was employed to rehearse the procedures and test assessment application (Lime Survey). Three pilot sessions took place on April 8th, 9th, and 12th, involving 46 students from three different schools.

During the pilot study, one minor problem with LimeSurvey options was identified – the setting blocking access to assessment for 5 minutes after unsuccessful attempts to log in was disabled as it caused too much trouble for interviewers at schools. Problems with inaccurate logging in are quite common in school computer-based studies. Apart from that, the pilot study did not identify any serious problems in the procedure nor in the assessments themselves, so the main data collection commenced as planned. The whole pilot administration was conducted in village schools because of the expected higher probability of technical problems, due to worse infrastructure of the school computer labs and lower capacity of the internet connection in rural areas. Overview of the pilot study outcomes are presented in Annex 3

9. Main Study

After conducting the pilot studies, during which data and paradata collection as well as fieldwork procedures were tested, the main study was carried out. Detailed fieldwork flow for the main study is presented in Table A3.3. and Figure A3.1 in Annex 3. Company's employees contacted schools from different regions of Poland.

In order to exclude participants not following the test protocol we have developed a set of criteria for excluding observations. This set comprised:

- **Drop Off**. Survey was not completed.
- **Speeding.** Survey completion time below 10 minutes.
- **Left browser card or window.** Leaving the survey window during the survey. This event was identified by recording the `window: blur` event in the collected paradata.
- **Different browsers or computers.** The same login (token) used on different computers – the web surveying platform we used blocks another access to the survey using a given login only at the survey completion. Consequently, it was technically possible to enter the survey at the same time using distinct computers. The event caused errors in the data collection, as two respondents were overwriting each other's responses. Such an event was rare, but indeed it happened a few times during the whole study. Main reason for that was the interviewer's failure to comply with the study procedures. This event was

identified based on the strings returned when navigating to subsequent screens by the JavaScript function `navigator.userAgent()`.

- **Parallel tests,** giving reasons to suspect cheating/cooperative completion. In the data, separately for each session, we have employed paired interviews comparisons, in the search of very similar interviews. The criteria for identifying pairs of interviews suspected of cheating/cooperative completion are as follows – all of these must be met simultaneously:
    - The start time of completing the survey differs by no more than one minute.
    - The average difference in the times of moving to subsequent screens (i.e., for each screen of the survey, we calculate the absolute difference between the completion times for each respondent and then calculate the average of these values for all screens for the given pair of respondents) is less than 15 seconds.
    - At most 2 different answers (treating skipping a question as an additional type of response) in the survey part where the respondent was supposed to answer attentively.
    - At most 2 different answers (treating skipping a question as an additional type of response) in the survey part where the respondent was supposed to answer inattentively.
- **Attempt out of the scheduled session.** Starting the survey more than 10 minutes after the session onset.
    - The moment of transitioning to the instruction screen, on which respondents are held for 30 seconds, was considered as the start of the survey.
    - The time when the token was entered, causing the welcome screen to be displayed, does not count as the interview start.

The excluded records from each of the test sessions are summed up in Table A3.3 in Annex 3.

10. Details of paradata collection

Paradata was collected using the *logdataLimeSurvey* version 1.2 Java Script applet installed on the *LimeSurvey* web surveying platform. Description of the technical details and list of collected events is available in the applet's web page https://github.com/tzoltak/logdataLimeSurvey.

Maximum time of paradata recording was limited to 5 minutes during a single entry of a given survey screen. Time distribution of the event collection was set to 100 milliseconds. Collected data was preprocessed using the logLime R package version 0.3. (Żółtak, 2024).

Sometimes paradata collection is unsuccessful due to technical errors, e.g. changing screen layout can break paradata collection on a given screen. Unfortunately, with the web surveying platform we used we had no technical ability to prevent such respodnets' behaviour. Table 2 summarizes the occurrence of such errors in the main study. Overall, despite that some errors were encountered in 17.1% of participants such problems eliminated only 2.9% out of the 36 702 collected paradata records.

Table 2

Proportion of Paradata Collection Errors

Problems with para-data collection	Overall		Attentive first, set one first		Attentive first, set two first		Inattentive first, set one first		Inattentive first, set two first	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
No problems	951	82.9%	234	80.4%	221	85.0%	248	82.7%	248	83.8%
Problem on at least one screen	196	17.1%	57	19.6%	39	15.0%	52	17.3%	48	16.2%
Overall	1147	100.0 %	291	100.0 %	260	100.0 %	300	100.0 %	296	100.0 %

11. Sample characteristics

Quota-based sampling was used to control for key socio-demographic characteristics: participants' gender and school location size. Target quotas were modeled after the Polish population distribution, as reported by the Polish Central Examination Commission for high-stakes school exams. Tables 3 and 4 display sample characteristics with regard to these two variables.

Table 3

Gender Distribution in the Sample (Main Study)

Gender	Number	Percent	Percent of valid answers	Target
Male	534	46.6%	50.4%	50.0%
Female	526	45.9%	49.6%	50.0%

Don’t want to answer	87	7.6%	-	-
Total	1147	100.0%	100.0%	100.0%

Note: Participants were provided with an option to refuse to answer the gender question. The option was provided on a request by the Ethical Research Committee.

Table 4

School Location Size Distribution in the Sample (Main Study)

Location size	Number	Percent	Target
Village	366	31.9%	35.0%
Town	470	41.0%	40.0%
City	311	27.1%	25.0%
Total	1147	100.0%	100.0%

Note: Village - pupils from schools located in villages and small towns of up to 5 thousand inhabitants, Town - pupils from schools located in towns of 5 - 100 thousand inhabitants, City - pupils from schools located in cities with more than 100 inhabitants.

The company was free to contact any school in Poland (non-probability quota-based sampling), but had to follow certain requirements:

- *Target population was set as students of grades VII and VIII of primary school.*
- *Gender distribution: 50% girls, 50% boys.*
- *The participants of the main study must come from at least 20 different schools.*
- *The schools must come from locations of various sizes, so as to reflect the composition of the Polish general population: rural schools (villages and towns up to 5 thousand inhabitants): 35%, schools from small towns (5-100 thousand inhabitants): 40%, schools from medium and large cities (over 100 thousand inhabitants): 25%.*
- *Schools can come from any region of Poland.*
- *A maximum of 4 classes from one school may participate in the main study.*

Among the respondents, 71.1% were from grade 7 and the remaining 28.9% were from grade 8.

The interviewers were instructed to exclude students who were not able to perform on the basis of their physical or intellectual disabilities, or due to being non-native language speakers. Final decision on student-level exclusions was upon the school headperson and supervising teachers.

As this study used a non-probability sampling, we have set minimum requirements on school and student sample size, which was set to 20 and 1000, respectively.

12. Datasets

In this section, we present how we constructed the datasets used in the analyses. This section also includes detailed codebooks that can be utilized during data inspection. The initial datasets were provided along with this report.

12.1 Defining AAoIs

Considering the layout of an item on the screen, we have decided to divide the screen area into three different areas of interest, each describing a functionally different part of the item (see Figure 3 below):

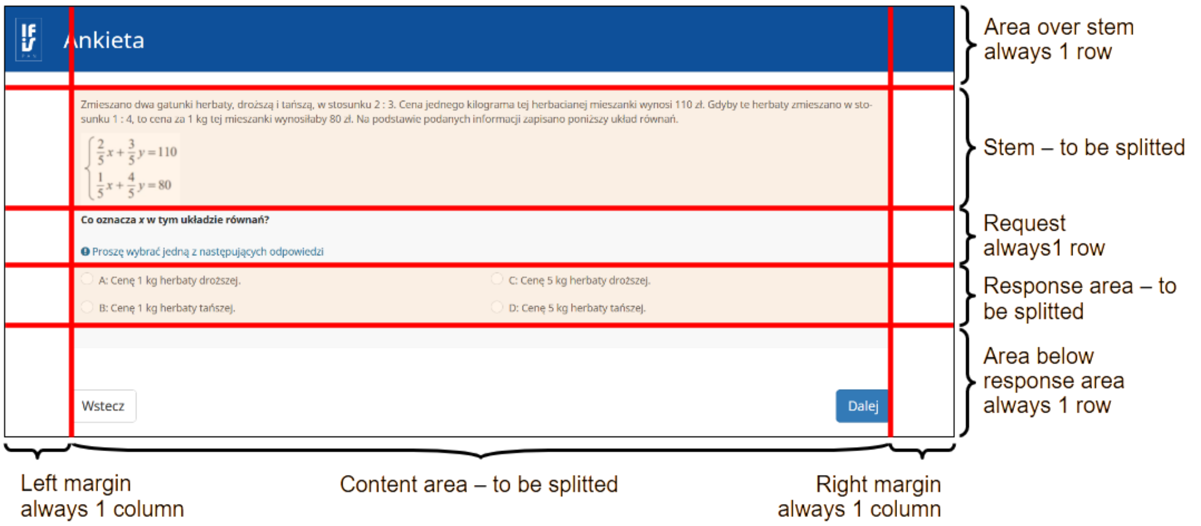
1) **Stem**, containing a description of the problem to solve,

2) **Request (lead-in)**, containing the final question asked or actions demanded from the examinee to solve the item,

3) **Responses**, containing a list of response categories to choose from.

Figure 3

AAoIs in Item - Example with one Split of Stem and Responses Areas



The simplest grid of AAoIs was defined by dividing the screen at the boundaries of these areas, resulting in five rows and three columns. More complex grids were created by applying evenly spaced splits to the stem and response areas in both horizontal and vertical directions. For example, Figure 4 illustrates AAoIs grid as a result of one vertical and one horizontal split of the stem and response areas (both highlighted in Figure 4), resulting in splitting each of these areas into four different AAoIs (cells): 6, 7, 10, and 11 (stem) and 18, 19, 22, and 23 (responses). Please note that other, less critical parts of the screen were also divided by extending these lines to the limits of the screen. Thus, the top of the screen with just a survey header and our Institute's logo is also split in four cells (1, 2, 3, 4), the area on the sides of the item stem as well (areas 5, 8, 9, 12), as well as area on the sides of the response area (17, 20, 21, 24) and the area below it (25, 26, 27, 28). The remaining four cells belong to the request (lead-in) area containing the actual question being asked (areas 13, 14, 15, 16).

Generally, the number of columns will be three (one for each: left margin, content area and right margin), plus the number of splits. In turn the number of rows in the grid equals five (one for each of the: area above stem, stem, request (lead-in), responses and area below responses) plus two times the number of splits (as both stem and response areas are splitted). Therefore, the number of AAoIs, as a function of the number of splits denoted by *s*, is given by: (s + 3)(5 + 2s). Scenarios with the number of splits varying from zero to six were considered in our study. See Table 5 for additional information.

Figure 4

AAoIs in Item - Example with one Split of Stem and Responses Areas

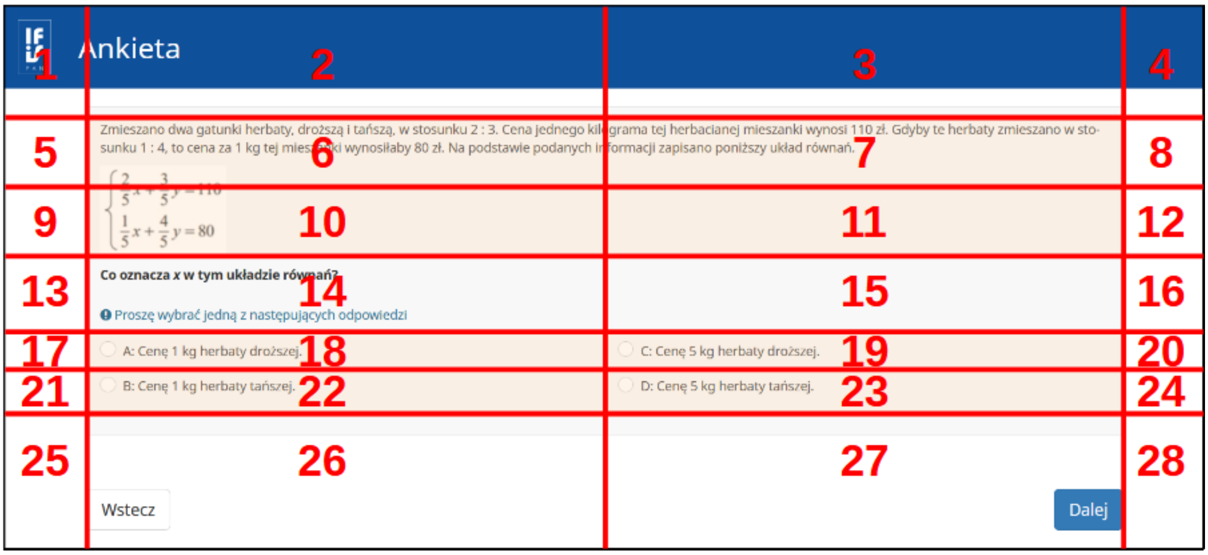


Table 5

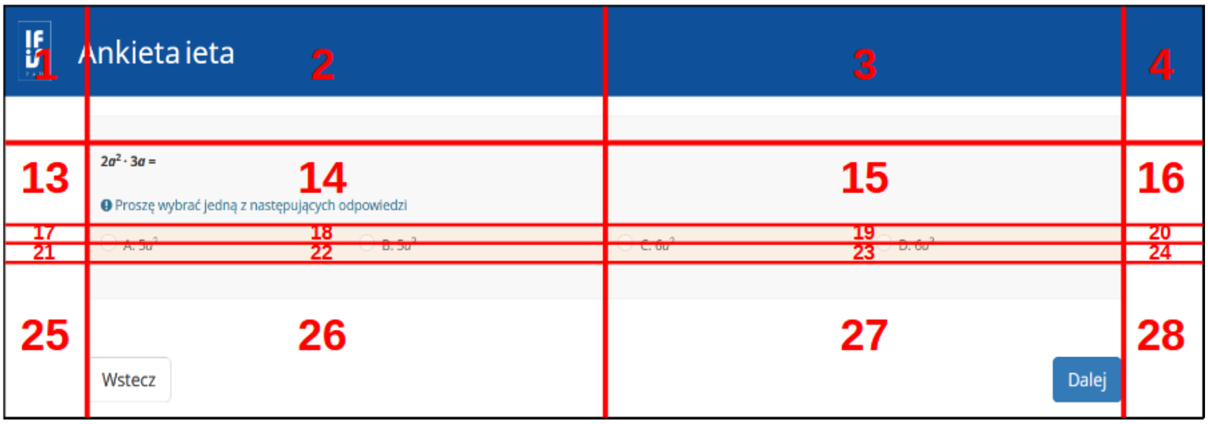
Number of AAoIs as the Function of Number of Splits

Number of splits of *stem* and *responses* areas	Number of AAoIs
0	15
1	28
2	45
3	66
4	91
5	120
6	153

As few items used in the test were very brief, it was impossible to distinguish the stem area in all items. In such instances, we did not create AAoIs corresponding to the stem area, while labeling the remaining AAoIs in a way preserving their functional meaning. Figure 5 illustrates the grid for such a "stemless" item for one split condition: AAoIs covering the area above the item content have labels 1 to 4; there are no AAoIs labeled 5 to 12, as there is no stem area (and its splits). AAoIs defined by the position of the response area are labeled 17 to 24 and AAoIs covering the area below the response options are labeled 25 to 28, analogously as in Figure 4. Areas from 13 to 16 code the request (lead-in) item part.

Figure 5

AAoIs in Item - Example with one Split of Responses Area and no Stem area



12.2 Datasets with AAoI information

We were able to use data from 1213 test-takers – including respondents from both pilot administration and the main study – covering 39,030 different item-respondents. Due to technical reasons we were forced to exclude 2.9% of all item-respondents. These reasons included: broken paradata records, leaving the assessment web browser tab while responding to a given item, resizing the web browser tab while responding to a given item or before returning to a given item.

Datasets describing test-takers actions (mouse moves and clicks) were prepared in four different variants (formats) defined by the crossing of two criteria:

1. Datasets describing separate respondent-items (called *long* format below) vs. describing respondents (called *wide* format below).

2. Datasets describing respondent clicks and moments when mouse cursor crosses boundaries between AAoIs vs. describing cursor position (as defined by AAoIs) at the end of each second.

More information on how the datasets were prepared and labeled is available in Annex 5 (see at the end of the report).

13. Manipulation check and exclusion criteria

To ensure comprehensive understanding of the instructions, interviewers provided an in-depth explanation of the study's guidelines and outlined the tasks participants were expected to complete. Participants were instructed to carefully read the instructions and subsequently answer a manipulation check question designed to assess their understanding:

"Before we begin, we want to ensure you understand the instructions. How do you plan to approach these tasks?"

Options provided were as follows:

1. *"I should answer the questions without much effort; I can guess."*
2. *"I should solve the tasks as best as I can."*
3. *"I should read them, but I am unsure what to do."*
4. *"I should ignore the instructions and act according to my judgment."*

Participants who selected an incorrect response were redirected to review the instructions and complete the manipulation check again. This iterative process continued until a correct response was provided, maximizing understanding of task expectations before the study's commencement.

However, despite these preliminary measures, some participants did not consistently follow instructions throughout the assessment. As a result, a post-hoc data cleaning process was implemented to exclude respondents who failed to meet the study's criteria for engagement and motivation, thus preserving the integrity of the study's findings.

- High-Motivation Condition: In the high-motivation condition, responses from participants who indicated minimal effort were excluded. Specifically, if participants selected option "d" ("I did not try at all") or “c” (“I put in little effort”) in response to the question ("How much effort did you put into accurately completing the test as instructed?"), they were removed from further analysis. This exclusion criterion ensured that only data from participants who reported at least a baseline level of engagement were retained. The application of this criterion resulted in the exclusion of 183 observations (15.96% of the total sample).
- Low-Motivation Condition: In the low-motivation condition, participants who selected option "d" ("I carefully read the questions and thoroughly analyzed all available answers before selecting one") to the question, "What was your strategy for answering the questions where we asked you to guess?" were excluded. This criterion was applied to differentiate genuinely low-motivation responses from those who over-engaged in the test despite the instructions. This process led to the exclusion of 287 observations (25.02% of the total sample).

These targeted exclusions allowed to retain data aligned with each condition's instruction requirements, ensuring that the sample contains only participants who acted in accord with the experimental manipulations. The whole process is summed up in Table 6.

Table 6

Number of Excluded Observations

Criterion	**Item set one**	**Item set two**	**Overall**
Reporting no motivation in the regular condition	94	89	183
Reporting attentive answering in the decreased motivation condition	145	142	287
Overall	239	231	470

Note: There were 13 participants who fulfilled both criteria (i.e. one criterion regarding item set one and another criterion regarding item set two).

Following the application of the exclusion criteria, the manipulation effect is observable in terms of the difference in sum score and overall time between the experimental conditions. The sum score reflects participants' overall task performance, while overall response time provides insight into the level of engagement between conditions. The data analysis reveals a significant difference between the attentive and lowered motivation groups (for score: F=51.03;   Prob > F = 0.0000; for time: F=220.52;   Prob > F = 0.0000). Detailed breakdown of item-level performance, are presented in the following section. Below, we summarize the main outcomes in both a table and a graphical representation.

Table 7

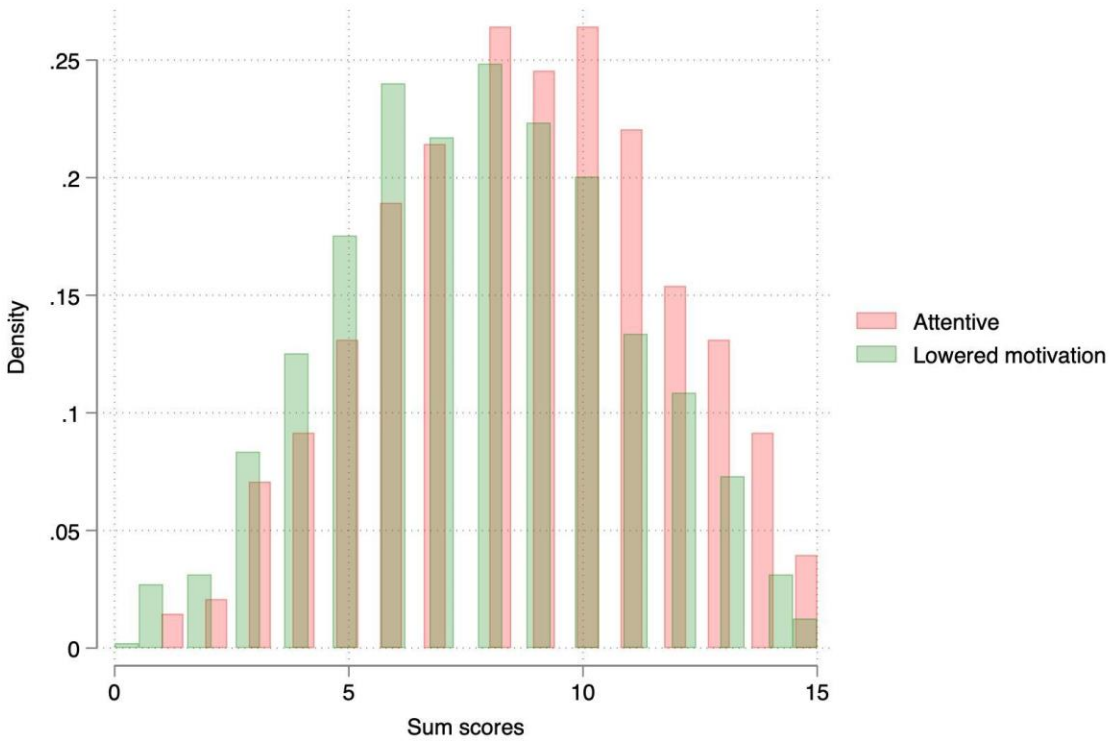
Sum Score and Overall Test Time Comparison between Experimental Conditions

	Attentive		**Lowered motivation**		**Mean difference**	
	Mean	**SD**	**Mean**	**SD**	**Score**	**SMD**
Sum score (0-15)	8.7	3.0	7.8	3.0	1.0	0.3
Test time (minutes)	10.4	3.8	8.0	3.2	2.4	0.7

Table 7 shows that the respondents in the attentive condition achieved a higher sum score (M = 8.7, SD = 3.0) compared to the lowered motivation condition (M = 7.8, SD = 3.0), with a mean difference of 1.0 and a standardized mean difference (SMD) of 0.3. Similarly, overall test time was longer in the attentive condition (M = 10.4 minutes, SD = 3.8) than in the lowered motivation condition (M = 8.0 minutes, SD = 3.2), resulting in a mean difference of 2.4 minutes and an SMD of 0.7. This shows that in the attentive condition the participants solved more items correctly and devoted more time to solve them.

Figure 6

Score Distribution between Experimental Conditions

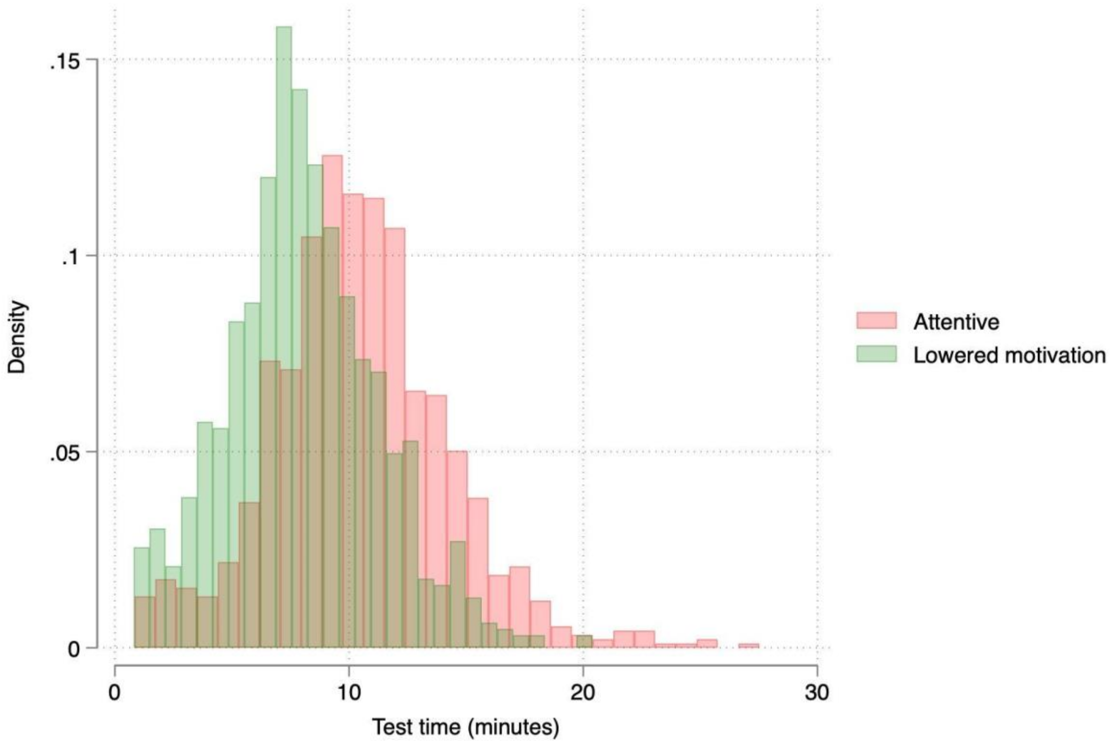


While interpreting the results, it is challenging to definitively assess the significance of an effect size of 0.3 SD, though it may be considered moderate. On one hand, studies such as the Trends in International Mathematics and Science Study (TIMSS) often report larger differences between countries. However, in the entire history of TIMSS, no single country has recorded either an increase or decrease in scores between the TIMSS cycles reaching an effect size of 0.3 or higher. This context suggests that, while moderate, the observed effect in our study is meaningful and may reflect a substantial shift in engagement and performance within the conditions established by our manipulation.

To further put this results in a context, in a research conducted by Liu and colleagues (2015), where, in a between-subjects design study, an instruction manipulation aimed at increasing participants' motivation led to an increase of test scores by 0.63 standard deviations (but only by 0.23 standard deviations when inattentive participants were removed based on time analysis). This study also showed that students in the motivated condition spent more time solving test items - similarly as in our study.

Figure 7

Overall Test Time Distribution between Experimental Conditions



In terms of response time, a difference of one standard deviation is substantial and can be interpreted as a relatively large effect. This magnitude indicates that the time spent by participants in the attentive condition was considerably longer compared to those in the lowered motivation condition, underscoring a significant behavioral shift in engagement. As previously evidenced by e.g. Arslan and Finn (2023), response processes, including response time, are often more affected by instruction manipulations than test performance. One of the reasons is that even rapid guessers can guess correctly quite often, e.g. in the case of easy multiple-choice items (Michaelides & Ivanova, 2022). Similar results were also obtained in our study (see Figure 7).

14. Psychometric analysis of the assessment

14.1 Data preparation

To ensure result interpretability, we excluded item scores where students likely did not follow experimental instructions. As described earlier, criteria for detection included: reporting attentive responses despite instructions to respond inattentively, and reporting "no motivation at all" while in the regular (attentive) condition.

Exclusions were done by substituting scores of the items in the appropriate part of the test for a given participant by missing data before performing psychometric analysis. Number of exclusions is summarized in Table 6. Two separate variables were created to store answers to the faulty item: one describing its faulty variant and the other describing its correct variant with students who did not take a given version of the item having missing value assigned as its score.

14.2 Model specifications

Psychometric properties of the assessment were studied using a set of two-group, unidimensional 2PL IRT[1] models in which group membership was determined by the set of items which the student took in the regular (attentive) condition. As the assignment to these groups was randomized and item set order was counterbalanced across the sample, we assumed the same, standard normal, distribution of the ability in all models. Consequently, potential differences between experimental conditions should manifest in differences between item parameters – it may be expected that items taken under the induced inattentiveness should have higher difficulty and lower discrimination in comparison to items taken in attentive condition.

We estimated three models differing in the specification of potential between-group differences of item parameters:

1. Full-invariant 2PL model - assuming no differences in item parameters between the experimental conditions.
2. Discrimination-invariant 2PL model - assuming no differences in discrimination parameter, but allowing differences in difficulty parameter.
3. Noninvariant 2PL model - allowing items to differ between the conditions in terms of both discriminancies and difficulties.

Model parameters were estimated using R package *mirt* (Chalmers, 2012), version 1.42.

[1] We considered also using a 3 PL IRT model, including the pseudo-guessing parameter, but with most items being relatively easy and limited number of available observations (especially in model specification allowing item parameters to vary between conditions) it provided very imprecise estimates, what led us to the decision of restraining to the 2 PL model.

14.3 Model fit comparison

Table 7 shows fit statistics for different 2PL model specifications.

Table 7

Model Fit Statistics for Different Model Specifications

Model	logLik	AIC	BIC	SABIC	LR test against more restrictive model		
					χ^2	df	p
Full-invariant 2PL	**−16,840.5**	33,809.1	**34,134.8**	33,931.5			
Discrimination-invariant 2PL	−16,757.5	**33,702.9**	34,181.4	**33,882.8**	166.1	30	0.000
Noninvariant 2PL	−16,735.1	33,718.2	34,349.4	33,955.5	44.7	30	0.041

Note: Analysis contains both pilot and main study samples.The smallest value of each information criteria is bolded.

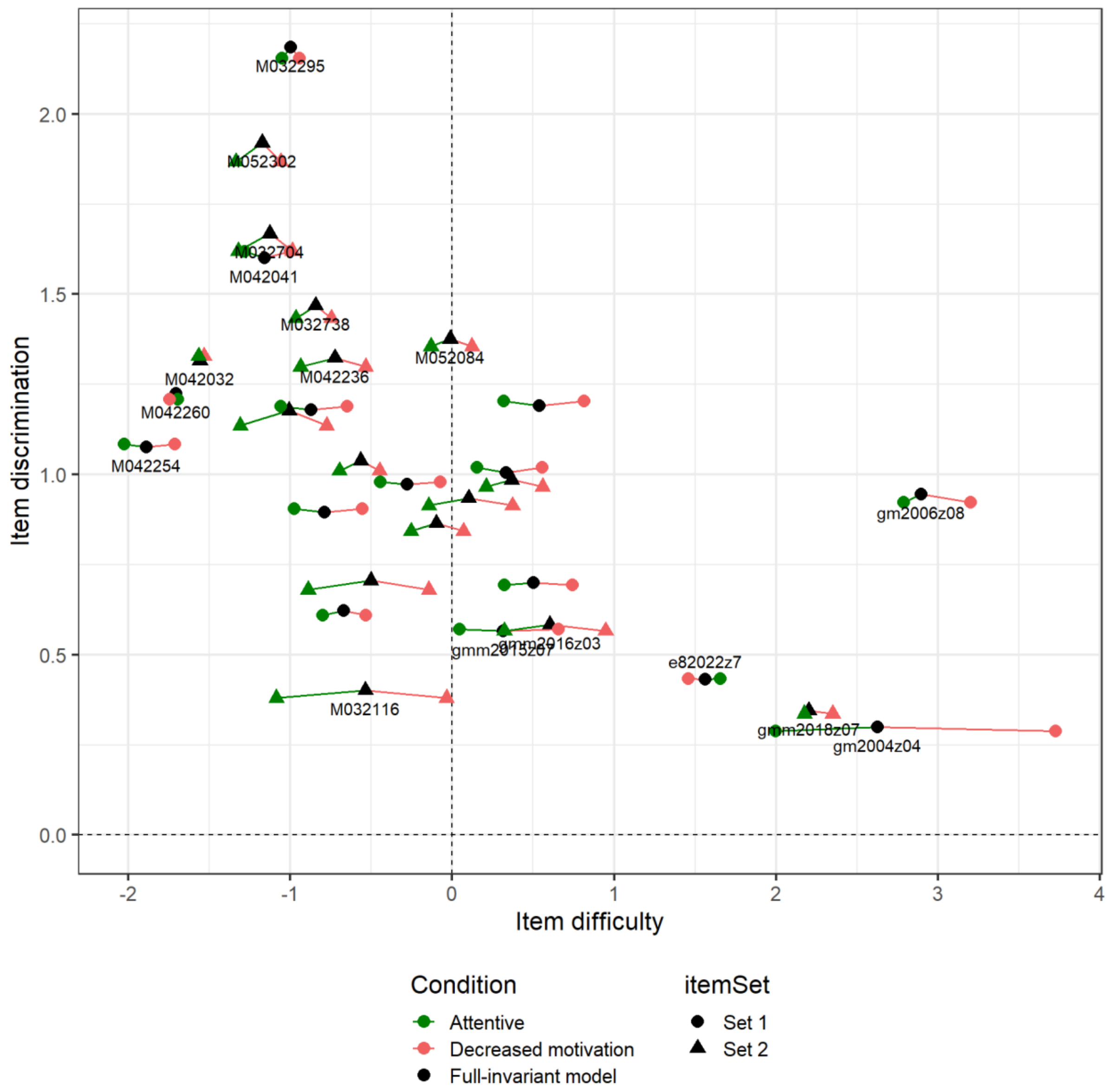
The full-invariant model has the lowest log-likelihood and BIC, but the highest AIC values. On the other hand, AIC and SABIC favor the discrimination-invariant model.

The results are somewhat ambiguous, yet, given AIC and SABIC values, we decided to interpret the data under the discrimination-invariant model. This suggests that, while the factor structure (as indicated by item discrimination, showing how items relate to the measured construct) remained stable even under reduced motivation, item difficulty shifted between the conditions.

14.4 Item parameter differences between conditions

Figure 8

Change of Item Parameters between Model Specifications



Note: Analysis contains both pilot and main study samples.

Both Figure 8 and Table 8 present similar information, suggesting that item difficulty generally increases under reduced motivation conditions, as seen in items such as M032579, where difficulty shifts from -1.00 in the attentive condition to -0.65 in the reduced motivation condition. However, some items remain relatively stable in their difficulty across conditions, like item M032295, which shows nearly identical values (-0.96 in the attentive condition and -0.95 in reduced motivation). The extent of difficulty variation between conditions also differs significantly by item; for instance, gm2004z04 shows a large shift in difficulty (from 2.24 to 3.70), while M042254 has a smaller change (from -1.93 to -1.66). These patterns indicate that the size of difficulty shift may vary, with some items being more affected by the reduced motivation condition than others.

Previous studies (e.g. Goldhammer et al., 2017; Michaelides & Ivanova, 2022) described such item characteristics as length, difficulty, presence of graphical material, and domain as related to propensity of rapid guessing and lower examinee effort.

Table 8

Item Parameters in Different 2PL Model Specifications

Item	Full-invariant 2PL model parameters		Slope-invariant 2PL model parameters			Differences in answer distributions (attentive vs. lowered motivation): p-value (Holm's correction)
	Discrimination	Difficulty	Discrimination	Difficulty - attentive	Difficulty - lowered motivation	
Item set 1						
M032579	1.18	−0.87	1.19	−1.06	−0.65	0.332
M042254	1.08	−1.89	1.08	−2.02	−1.71	1.000
M042222	0.97	−0.28	0.98	−0.44	−0.07	0.920
gm2006z08	0.95	2.90	0.92	2.79	3.20	0.236
M042260	1.22	−1.71	1.21	−1.69	−1.74	1.000
gmm2015z07	0.57	0.32	0.57	0.04	0.66	0.432
M032295	2.18	−1.00	2.16	−1.05	−0.94	1.000
M042031	1.19	0.54	1.20	0.32	0.82	**0.005**
gm2004z04	0.30	2.63	0.29	2.00	3.73	0.067
M032679	0.62	−0.67	0.61	−0.80	−0.53	1.000
M032198	0.70	0.50	0.69	0.32	0.74	0.920
gmm2014z11	1.01	0.33	1.02	0.15	0.56	0.529
M042041	1.60	−1.16	1.62	−1.28	−1.00	1.000
e82022z7	0.43	1.56	0.43	1.66	1.46	0.920
M032166	0.90	−0.79	0.90	−0.98	−0.55	0.920
Item set 2						
M042032	1.32	−1.56	1.33	−1.57	−1.53	1.000
M052302	1.92	−1.17	1.87	−1.34	−1.06	**0.004**
gmm2016z20	0.71	−0.50	0.68	−0.89	−0.14	**0.017**
M032738	1.47	−0.84	1.43	−0.97	−0.75	0.225
M032477	0.93	0.10	0.91	−0.14	0.37	**0.029**
gmm2018z07	0.34	2.20	0.34	2.17	2.35	1.000
M032704	1.67	−1.13	1.62	−1.32	−0.99	0.067

Item	Full-invariant 2PL model parameters		Slope-invariant 2PL model parameters			Differences in answer distributions (attentive vs. lowered motivation): p-value (Holm's correction)
	Discrimination	Difficulty	Discrimination	Difficulty - attentive	Difficulty - lowered motivation	
M032116	0.40	−0.54	0.38	−1.08	−0.03	0.060
gmm2016z01	0.99	0.37	0.97	0.21	0.56	0.460
M042236	1.32	−0.72	1.30	−0.94	−0.53	**0.029**
gm2002z05	0.86	−0.10	0.84	−0.25	0.07	**0.001**
gmm2016z03	0.58	0.60	0.57	0.32	0.95	0.236
M022101	1.18	−1.01	1.14	−1.31	−0.77	**0.001**
M052084	1.38	−0.01	1.35	−0.13	0.12	0.528
M032094	1.04	−0.56	1.01	−0.70	−0.45	**0.029**
Faulty item						
Faulty variant	1.10	−3.66	1.07	−3.79		
Correct variant	1.04	−1.27	1.04	−1.29		

Notes: Analysis contains both pilot and main study samples. Both variants of the faulty item were held full-invariant also in the discrimination-invariant model. Differences in answer distributions were tested using Pearson's chi-square test of independence. P-values were corrected for multiple comparisons using Holm's correction. P-values lower or equal to 0.05 marked in bold.

14.5 Differences in item answers and response time distributions

Table 9 compares item-level response time differences across experimental conditions. In contrast to item difficulty parameters, most of the items' response times differ between the conditions (with the exception of e.g. M042222 or M032198), with items in the inattentive condition having faster response times.

Table 9

Differences in Item Response Times between Experimental Conditions

Item	Mean log(response time)		T-test for differences in means p-value (Holm's correction)
	Attentive condition	Decreased motivation condition	
Item set 1			
M032579	3.39	3.29	**0.035**
M042254	3.92	3.81	**0.035**
M042222	3.64	3.55	0.054
gm2006z08	3.76	3.61	**0.011**
M042260	3.43	3.32	**0.011**
gmm2015z07	3.92	3.63	**0.000**
M032295	3.28	3.13	**0.002**
M042031	3.56	3.32	**0.000**
gm2004z04	3.28	3.04	**0.000**
M032679	3.52	3.32	**0.000**
M032198	3.14	3.03	**0.026**
gmm2014z11	3.76	3.54	**0.000**
M042041	3.12	2.98	**0.014**
e82022z7	3.40	3.15	**0.000**
M032166	3.15	3.02	**0.027**
Item set 2			
M042032	2.61	2.54	0.054
M052302	3.23	3.03	**0.000**
gmm2016z20	3.69	3.36	**0.000**
M032738	3.07	2.92	**0.000**
M032477	3.79	3.33	**0.000**
gmm2018z07	3.69	3.16	**0.000**
M032704	3.69	3.30	**0.000**
M032116	3.39	2.92	**0.000**
gmm2016z01	4.09	3.56	**0.000**
M042236	3.26	2.94	**0.000**
gm2002z05	3.75	3.22	**0.000**
gmm2016z03	3.80	3.31	**0.000**
M022101	3.45	3.09	**0.000**
M052084	3.21	2.92	**0.000**
M032094	3.24	2.84	**0.000**

Notes: Analysis contains both pilot and main study samples.P-values were corrected for multiple comparisons using Holm's correction. P-values lower or equal to 0.05 marked in bold.

14.6 Faulty item

Faulty item variant was manipulated in the between-subjects design, in contrast to the motivation condition, which was manipulated in the within-subjects schema. This item always appeared at the very end of the motivated condition part of the assessment. The item appeared either in the correct version or in the flawed (faulty) version. The faulty version contained four correct response options (out of five presented), while the correct variant had only one correct response option (out of five presented).

Responses to the faulty item were modeled differently than other items - its parameters were fixed to constant across both experimental conditions in each of the compared model specifications (full-invariant and discrimination-invariant 2PL).

We have compared faulty item's item characteristic curves from 2 PL IRT model between the two conditions, pointing to large differences in item psychometric characteristics between the two conditions (Figure 9).

Figure 9

Item Characteristic Curves of the Faulty Item

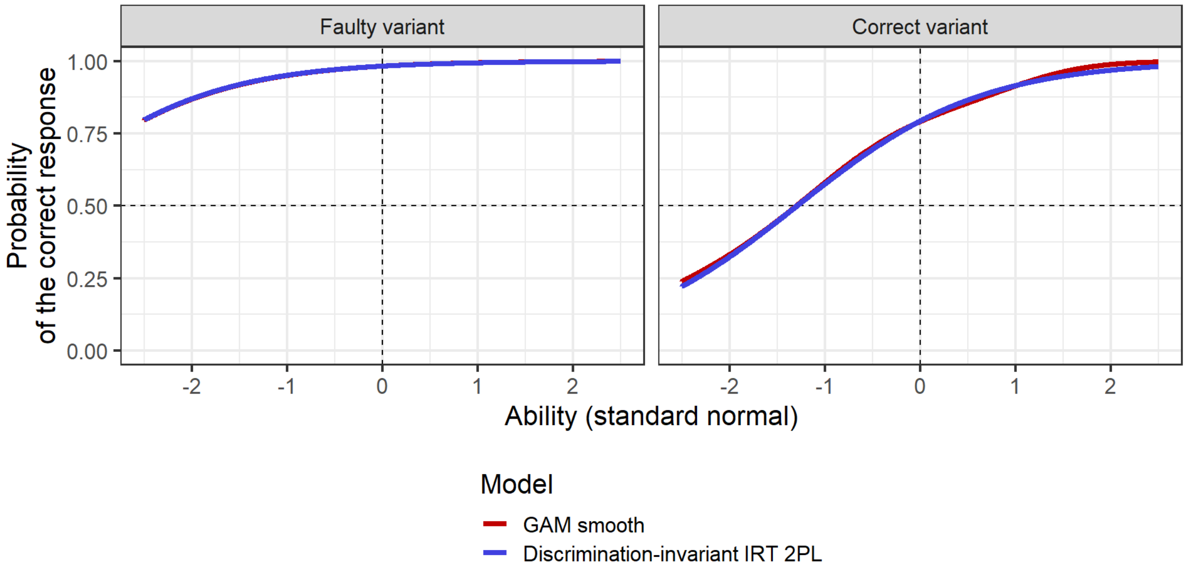


Table 10

Response Distribution to the Faulty Item

Variant	Answers					Overall
	A (correct in the faulty variant)	**B (correct in the faulty variant)**	**C (correct in the faulty variant)**	**D (correct in the faulty variant)**	**E (correct in the correct variant)**	
Faulty	270 (52.1%)	203 (39.2%)	21 (4.1%)	10 (1.9%)	14 (2.7%)	518 (100.0%)
Correct	27 (5.3%)	46 (9.0%)	29 (5.7%)	22 (4.3%)	388 (75.8%)	512 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(4) = 651.53$, $p = 0.000$

Note: Analysis contains both pilot and main study samples.

Figure 10

Faulty Item Response Time Distribution

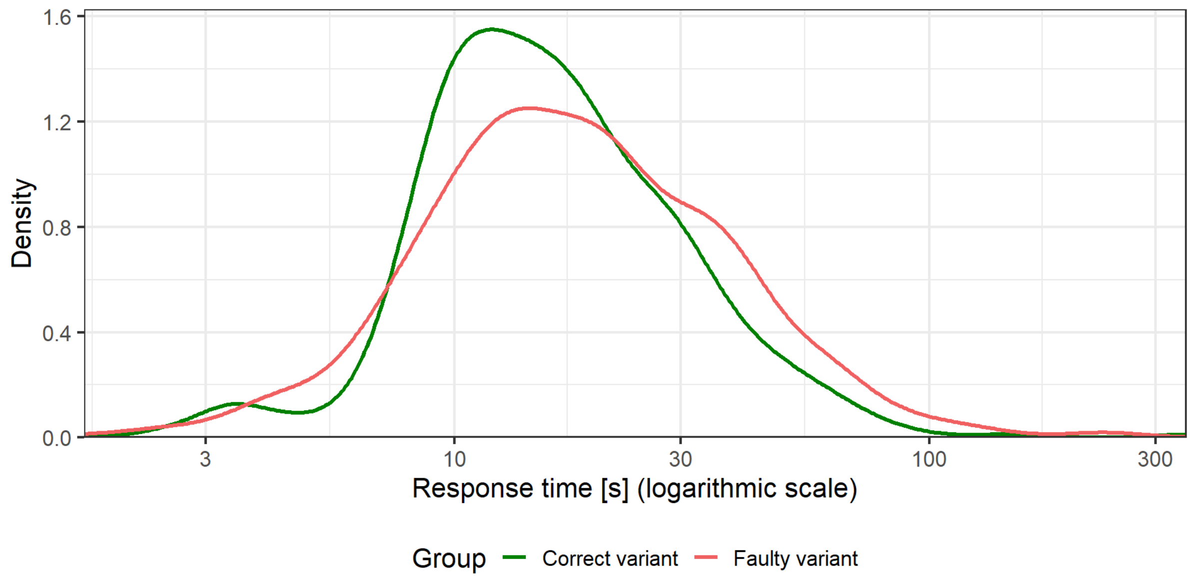


Table 11

Comparison of Mouse Movement Indices between Faulty Item Variants

Measure	Mean value		Cohen's d	*t* statistic	p-value	R^2
	Correct variant	Faulty variant				
log(dX)	7.83	8.09	0.44	7.59	**0.000**	0.046
log(dY)	6.70	6.79	0.14	2.39	**0.017**	0.005
sqrt(flipsX)	2.96	3.05	0.08	1.36	0.175	0.002
sqrt(flipsY)	3.36	3.54	0.14	2.40	**0.016**	0.005
log(vX)	−1.73	−1.56	0.26	4.57	**0.000**	0.017
log(vY)	−2.86	−2.86	−0.01	−0.15	0.883	0.000
log(aX)	−6.45	−6.30	0.23	3.88	**0.000**	0.013
log(aY)	−7.57	−7.56	0.00	0.06	0.948	0.000
log(response time)	2.76	2.85	0.14	2.46	**0.014**	0.005

Notes: N = 1189. Analysis contains both pilot and main study samples. P-values lower or equal to 0.05 marked in bold.

log(dX) - distance travelled horizontally, logged; log(dY) - distance travelled vertically, logged; sqrt(flipsX) - square root of the number of direction changes in horizontal moves; sqrt(flipsY) - in vertical moves; log(vX)/(vY) - speed of horizontal/vertical moves, logged; log(vX)/(vY) - acceleration of horizontal/vertical moves, logged.

In order to compare response processes between the two variants of the item we have used mouse move indicators that comprehensively sum up participants' behavior in various computer-based tasks. We constructed four sets of indicators based on cursor movement: (1) total distance traveled, (2) average speed, (3) average acceleration of cursor move, and (4) number of flips, that is, changes of direction of movement (cf. Horwitz et al., 2017).

For the total distance traveled we distinguished the horizontal direction (dX) and vertical direction (dY) in pixels (px). Prior research showed that the position of the cursor was relatively close to the position of the eye gaze on the screen (Clark & Stephane, 2018; Kirsh & Joy, 2020); therefore, we stipulated that the distance traveled reflected the survey response–related behaviors. A short distance traveled should therefore indicate less reading, less hesitation, and a smaller number of opinion changes or adjustments to the response categories, and, therefore, it could be related to more C/IER. The shortest possible distance traveled indicated a straight line, and, therefore, it seemed natural that the distance traveled should be a good measure of straightlining. We assumed that the horizontal dimension was of more importance than the vertical one because reading (in Indo-European languages) is based on the horizontal dimension, from left to right. Indeed, many participants treated the cursor as a reading aid (Horwitz et al., 2017). However, the vertical dimension should also be of some importance, because a long distance traveled in it may indicate going back to items and changing the responses, or thinking of changing them, and this could indicate more careful responding. There was also evidence that an excessive distance traveled could indicate confusion and difficulty in making a response (Stieger & Reips, 2010).

Another two indices reported that the average speed of the cursor moves in horizontal (vX) and vertical (vY) directions (in pixels per second [px/s]). Following the theory that the cursor is relatively close to the position of the eye gaze on the screen, a high speed in both dimensions should indicate inattentiveness or speeding. The vertical speed could be a good indicator of straightlining, as we expected that it was related to rapid response patterns that follow a straight line.

The third set of indices reported the average acceleration of cursor moves in the horizontal direction (aX) and vertical direction (aY) (in pixels per squared second [px/s2]). High average acceleration could indicate distractions (Brenner & Smeets, 2003) or different emotional reactions in one's moves (Yamauchi & Xiao, 2018).

Finally, we counted the number of flips, that is, the changes in direction of movements, using the horizontal direction (flipX) and vertical (flipY) direction. On the one hand, we assumed that when respondents were moving the cursor across a computer screen toward a target, they used direction changes to reach the target for a mindful response. On the other hand, straightlining should be indicated by a rather small number of flips. A large number of flips could indicate hesitancy or indecisiveness (Horwitz et al., 2017).

See Pokropek et al. (2023) for a detailed description of the mouse moves indicators used.

The results of the comparison between correct and faulty versions of the flawed item indicated that:

- Respondents spent more time solving the item in the flawed version in comparison to the correct version (see Table 11 and Figure 10).
- Respondents mostly selected one of the two first correct response options in the flawed variant (see Table 10).
- Respondents traveled longer distances and moved their mouse with higher velocity and acceleration in the horizontal dimension in the faulty in comparison to the correct version (cf. Table 11).
- Respondents noted more vertical flips (changes of move direction) in the faulty vs. correct variant. To some extent respondents also traveled longer distances in the vertical direction in the faulty version of this item (cf. Table 11).

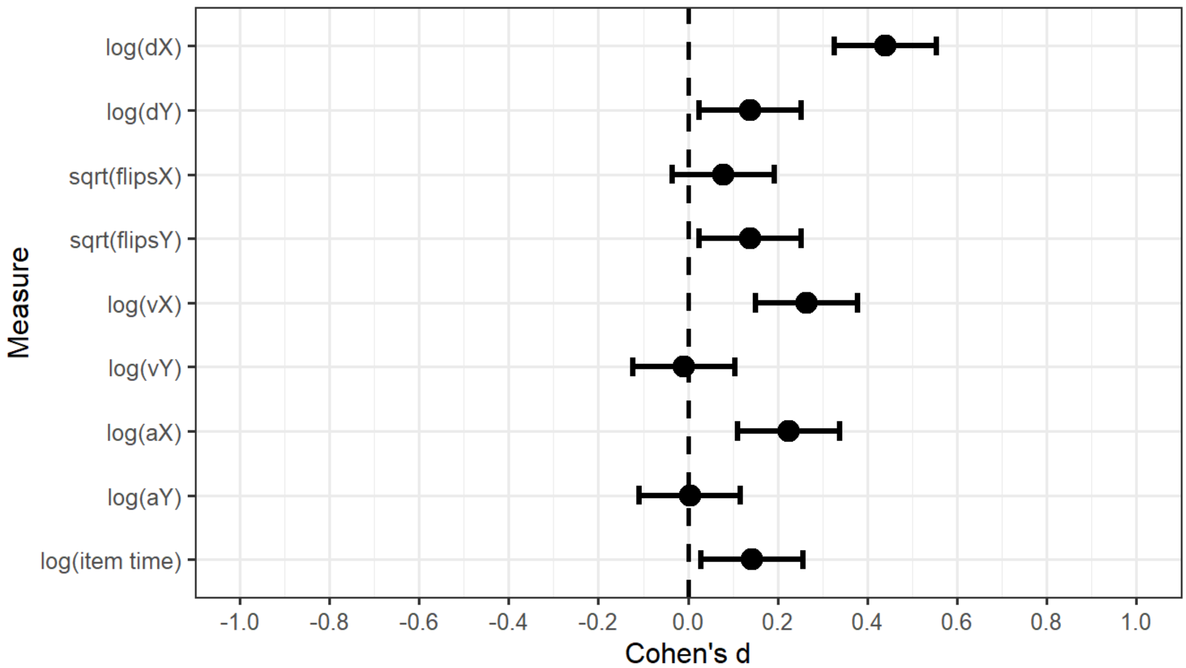
See Figure 11 for a graphical representation of the results in more detail.

These results indicate that the faulty item provided some amount of confusion and caused participants to think more on solving this item. Previous research (e.g. Horwitz et al., 2017, 2020) related subjective task difficulty to more mouse moves, especially horizontal traveling and vertical flips, just as in our study. Leipold et al. (2024) linked a high number of vertical flips to hesitation which option should be selected in items with very similar response options.

Our result corroborates previous results and points out to usefulness of paradata, with an example of mouse moves, in screening measurement tools for design flaws also in the context of cognitive tests, especially as other similar studies concentrated on surveys and other cognitive tasks (e.g. Fernandez-Fontelo et al., 2023; Meidenbauer et al., 2023).

Figure 11

Differences in Mouse Movement Indices between Faulty Item Versions



Notes: Values for Mouse Movement Indices in the Faulty Version are compared to the values in the Correct Version. Analysis contains both pilot and main study samples. Bars' widths represent 95% confidence intervals. Analysis contains both pilot and main study samples.

15. Using AAoIs for detecting lowered motivation and flaws in items

15.1 General modeling approach

The proposed approach is based on a classification task that utilizes sequential data from students' interactions with items during computer-based assessments. The classification of sequential data is a critical task in various domains, including natural language processing, time-series analysis, and behavioral studies (Graves, 2012). Long Short-Term Memory (LSTM) neural networks have been widely adopted due to their ability to capture long-term dependencies in sequential data (Hochreiter & Schmidhuber, 1997). However, optimizing LSTM models requires careful tuning of hyperparameters to balance model complexity and generalization performance (Goodfellow et al., 2016).

While Long Short-Term Memory (LSTM) networks were chosen for this study due to their proven ability to handle sequential data, other neural network architectures could have been considered for the task. One such alternative is the Gated Recurrent Unit (GRU), which, like LSTM, is designed to capture dependencies in sequential data. GRUs have a simpler architecture, with fewer parameters than LSTMs, making them potentially more efficient for certain applications while maintaining comparable performance. Pokropek, Żółtak, and Muszyński (2024) explored both GRU and BLSTM architectures in a similar context of analyzing sequential data. Their results showed no substantial differences in performance between GRU and LSTM, but they observed a slight advantage for BLSTM due to its ability to utilize information from the full sequence more effectively

In this analysis, we explored several different approaches, each representing various implementations and adaptations of the Approximate Areas of Interest (AAOI) method. First, we investigated different grid densities, resulting in 15, 28, 60, and 120 segments. Although we initially created more variations, preliminary analyses indicated that there were no significant differences between these grid densities.

Additionally, we built predictive models using varying test lengths (measured by the number of items) to predict motivation. Specifically, we utilized information from 15 items (representing the full test for one condition), 5 items, and 3 items to assess model performance with different amounts of information. This allowed us to test the robustness of the model under varying data conditions, but also pointed to the importance of the amount of item-level information.

Finally, we explored additional estimation techniques based on data augmentation, with the expectation that this would enhance the predictive power of our model. These augmentation methods aimed to enrich the training data and potentially improve the model's generalization to unseen cases.

For each data type in the first step we use Bayesian optimization for hyperparameter tuning, we seek to enhance model performance while efficiently exploring the hyperparameter space (Snoek et al., 2012). We focused on optimizing the initial learning rate and dropout rate, which are crucial for training stability and preventing overfitting (Srivastava et al., 2014).

The whole available data were divided into training, validation, and test sets. Training and validation sets were used for hyperparameter tuning, while in the second step optimized parameters were used to train final models on training data and their performance was evaluated using test data.

15.2 Data Preparation

The dataset comprised sequential data collected from participants, along with auxiliary information stored in separate CSV files. The sequences represented time-series data of participant interactions, and the auxiliary data included labels and additional features pertinent to the classification task. Participants with inconsistencies in their data (e.g., mismatched or incomplete information as detailed in the data section) were removed entirely from the analysis to prevent errors caused by unreliable or incomplete data. Sequences with fewer than three, five or 15 (depending on analysis condition) logged items were removed to ensure that the model was trained on full data. The filtered dataset was randomly shuffled and split into training (60%), validation (20%), and test sets (20%) (Goodfellow et al., 2016).

15.3 Model Architecture

An LSTM neural network was implemented to leverage its strength in capturing temporal dependencies within sequential data (Hochreiter & Schmidhuber, 1997). The model architecture consisted of:

- **Sequence Input Layer**: Configured to match the dimensionality of the input features.
- **Bidirectional LSTM Layers**: The network included a bidirectional LSTM with 128 units, processing input sequences while retaining information from both past and future states. A second LSTM layer with 64 units, configured with "last" output mode, focused on summarizing the sequence into a single output, enhancing the model’s contextual comprehension (Schuster & Paliwal, 1997).
- **Dropout Layers**: Dropout layers followed each LSTM and fully connected layer. The dropout rate, optimized via Bayesian techniques, regularizes the model by reducing overfitting and preventing neuron co-adaptation (Srivastava et al., 2014).

- **Fully Connected Layers**: A 32-unit fully connected layer mapped the learned representations to the output layer.
- **Softmax and Classification Layers**: These layers generated class probabilities and computed cross-entropy loss for training and evaluation.

This architecture balances sequence encoding with regularization, enabling robust classification on temporal data.

15.4 Hyperparameter Optimization and Model Training

Bayesian optimization was applied to tune hyperparameters, specifically the initial learning rate and dropout rate. Bayesian optimization efficiently searches the hyperparameter space by building a probabilistic model of the objective function and selecting hyperparameters to evaluate based on this model (Snoek et al., 2012). The optimization was limited to 10 evaluations to balance computational cost with search thoroughness.

The final model was trained using the best hyperparameters identified. Training was conducted using the Adam optimizer (Kingma & Ba, 2015). The training process included an initial learning rate that was optimized through Bayesian methods. The model was trained for a maximum of 50 epochs, with early stopping implemented after 4 validation epochs without improvement, in order to prevent overfitting. A mini-batch size of 256 was used to balance training speed and stability. To handle varying sequence lengths, sequences were either padded or truncated to the longest sequence length within each batch, ensuring computational efficiency.

15.5 Data augmentation attempts

One limitation of our study was the limited amount of training data available, which restricted our ability to achieve higher prediction accuracy. To address this, we tested data augmentation techniques commonly employed in such situations, including synthetic data generation methods designed to expand the training dataset by introducing slight variations into existing sequences. Data augmentation is a widely used approach to increase the diversity and volume of training data without requiring new samples, often resulting in improved model robustness and performance, especially when data collection is constrained (Shorten & Khoshgoftaar, 2019; Wen et al., 2020).

In our case, we implemented a MATLAB-based augmentation function designed specifically for binary sequences with a timestamp feature. This function performs augmentation by modifying sequences through random binary feature flipping, as well as inserting and deleting events in the sequences. The procedure begins by identifying the binary feature indices and timestamp feature index, ensuring consistency across sequences. Then, for each sequence, the function performs either an event insertion or deletion. For insertion, a random binary event is created, and a corresponding timestamp is generated based on the surrounding timestamps. For deletion, a random position is selected, and the event at that position is removed if the sequence remains above a minimum length threshold.

In addition to these transformations, a random binary feature flipping with a 5% probability was applied across events to introduce further variability. Synthetic data generation was used to double the size of the training dataset. Only sequences that retained a minimum number of time steps were included in the augmented dataset. This approach enabled us to expand the training dataset with slightly modified sequences, increasing the model's exposure to a broader range of scenarios while preserving the original data structure. This was applied only to training and validation data and not testing data

15.6 Model Evaluation

Performance was evaluated on the test set using several metrics. Accuracy was measured as the ratio of correctly predicted instances to the total number of instances. Precision, or Positive Predictive Value, was used to assess the proportion of true positives among all positive predictions, while recall, or True Positive Rate, measured the model's ability to identify all actual positives. Specificity, also known as True Negative Rate, evaluated the model's effectiveness in correctly identifying negatives. The F1 score, which is the harmonic mean of precision and recall, provided a balanced measure of the model's performance. Balanced accuracy, calculated as the average of recall and specificity, offered a more nuanced understanding of the model's ability to handle both positive and negative cases. Finally, baseline accuracy was determined by predicting the majority class, serving as a benchmark for comparison.

15.7 Predicting motivation

Table 12 presents performance metrics for models predicting the lowered motivation condition, with configurations varying by the number of items (sequence length) and grid dimensionality (number of the defined approximate areas of interest (AAoIs)). The grid dimension controls the resolution, with higher values increasing granularity and potentially enhancing model sensitivity by capturing finer behavioral nuances.

Table 12

Results of predicting motivation conditions using ML models.

N items	Grid	Accuracy	Precision	Recall (TPR)	F1 Score	Specificity	Balanced Accuracy
15	15	0.626	0.628	0.533	0.577	**0.711**	0.622
15	28	**0.673**	0.650	**0.848**	**0.736**	0.469	**0.659**
15	66	0.632	0.639	0.689	0.663	0.570	0.629
15	120	0.591	0.613	0.522	0.564	0.662	0.592
5	15	0.611	**0.681**	0.536	0.599	0.700	0.618
5	28	0.635	0.657	0.686	0.671	0.575	0.631
5	66	0.616	0.619	0.685	0.651	0.541	0.613
5	120	0.602	0.594	0.711	0.647	0.487	0.599
3	15	0.623	0.609	0.719	0.659	0.525	0.622
3	28	0.587	0.622	0.601	0.611	0.572	0.586
3	66	0.622	0.603	0.816	0.693	0.410	0.613
3	120	0.598	0.613	0.625	0.619	0.569	0.597

Note: Analysis contains both pilot and main study samples. The highest value in a given column is bolded.

Generally, increasing the sequence length improves prediction, as it provides the model with more comprehensive behavioral data associated with motivation. Notably, the configuration using 15 items and a 28-grid dimension yielded balanced metrics — Accuracy at 0.673, Recall at 0.650, and Balanced Accuracy at 0.659 — indicating that this medium-resolution grid setup with a longer sequence length captures motivational patterns effectively without overfitting.

As grid dimension increases (e.g., 66 and 120), performance shows diminishing returns. Although a higher resolution might theoretically capture finer details, limited training data limits this benefit, often resulting in overfitting. Overfitting is particularly evident in the declining values for Balanced Accuracy and Specificity at higher grid dimensions, likely due to sparse data per grid cell, which restricts the model's ability to generalize. This suggests that, although higher grid resolution can increase precision under ideal conditions, it may hinder performance when data points are limited.

Furthermore, models with fewer items (e.g., 5 or 3) tend to perform worse, affirming that longer sequences are critical for accurately capturing motivation-related behavior. However, certain shorter sequences (e.g., 3 items at a 66-grid) achieve moderate performance, with Recall at 0.816 and Balanced Accuracy at 0.613, likely due to increased grid detail capturing specific patterns.

In sum, a 15-item sequence with a 28-grid dimension best balances prediction accuracy and model generalizability in this dataset. Higher grid dimensions, although more detailed, often result in overfitting due to sparse data, whereas shorter sequences lack sufficient behavioral data for robust predictions.

Table 13 presents the results from predictive models utilizing the augmented data approach (described above), where we increased the data size by doubling the number of respondents. The computations were made for the condition with 15 items only, as it was the best performing scenario in the previous analyses.

Table 13

Results of Predicting Motivation Condition Using ML Models with Synthetic Data Approach

N items	Grid	Accuracy	Precision	Recall (TPR)	F1 Score	Specificity	Balanced Accuracy
15	15	0.616	0.619	0.604	0.611	0.629	0.616
15	28	**0.645**	**0.663**	0.690	**0.676**	0.592	**0.641**
15	66	0.610	0.561	**0.721**	0.631	0.515	0.618
15	120	0.616	0.628	0.581	0.604	**0.652**	0.617

Note: Analysis contains both pilot and main study samples. The highest value in a given column is bolded.

In our case, the synthetic data approach did not improve prediction outcomes. Once again, the model using a 28-grid resolution with data from 15 items yielded the best predictive accuracy. The subtleties of motivational differences in our dataset may not be sufficiently represented by augmented sequences, which are generated by minor modifications to existing patterns rather than new, truly diverse behavioral data. It is also worthy to note that response processes, as represented by mouse trajectories, mouse clicks, and corresponding timestamps is an innovative and relatively poorly understood type of research data and there is probably much to learn with regard to simulating such datasets.

15.8 Predicting flawed items

The prediction of whether a given item had deliberately introduced flaws was based on the same methodology and machine learning techniques used in detecting low motivation. However, in this analysis, we focused on a single item and aimed to predict if it was its flawed or correct version. The classification relied on analyzing the sequence of mouse movements and the Approximate Areas of Interest (AAoIs) data, collected while respondents struggled to solve this item. By applying a similar machine learning framework (the only change in these analyses was that the maximum number of epochs was set to 100 instead of 50), we used behavioral data (mouse movement patterns) and the AAI score to distinguish between flawed and correct versions of the item, enabling us to identify subtle differences in user interactions that may indicate the presence of faulty items in an assessment. In the analysis predicting whether an item was flawed, we used different Approximate Areas of Interest (AAoIs) resolutions, as defined by the number of grids (15, 28, 66, and 120). The grid count reflects the level of resolution in tracking user interactions, with higher grid numbers representing more detailed tracking. The results of these analyses are presented in Table 21.

Table 14

Results of Predicting Flawed Item Variant

N items	Grid	Accuracy	Precision	Recall (TPR)	F1 Score	Specificity	Balanced Accuracy
1	15	0.626	0.645	0.575	0.608	0.678	0.626
1	28	0.832	0.808	0.863	0.835	0.802	0.832
1	66	**0.861**	**0.825**	**0.904**	**0.863**	**0.821**	**0.863**
1	120	0.807	0.754	0.864	0.805	0.758	0.811

Note: Analysis contains both pilot and main study samples. The highest value in a given column is bolded.

At the lowest resolution of 15 grids, model performance was moderate, with an accuracy of 0.626, precision of 0.645, recall of 0.575, and an F1 Score of 0.608. This resolution appeared to limit the model's ability to capture nuanced interactions, resulting in only moderate classification effectiveness. Increasing the resolution to 28 grids led to significant improvements, with accuracy reaching 0.832, recall at 0.863, and F1 Score at 0.835. This improved tracking detail enabled the model to detect flawed items more effectively.

With 66 grids, performance was optimized, reaching the highest observed accuracy (0.861) and balanced accuracy (0.863), alongside a recall of 0.904 and F1 Score of 0.863. These results suggest that the model benefits from a high AAoI resolution in detecting the flawed version of the item, capturing subtle interaction differences with a refined spatial dimensionality.

At 120 grids, performance decreased slightly, with accuracy dropping to 0.807 and F1 Score to 0.805. In addition to potential noise introduced by extremely high resolution, it is also possible that the ratio of data quantity to resolution was a limiting factor. Given the model’s limited data availability, this high-resolution AAoI setting may have required a larger dataset to reach optimal performance.

Comparing these results with our previous prediction tasks, which focused on predicting motivation levels and achieved an accuracy below 68%, reveals that the method is effective across task types. While it excels in simpler tasks, such as detecting obviously flawed items, it also shows promise for more nuanced classifications, though with less pronounced success. Collectively, these results indicate that our approach performs well, with especially high accuracy in straightforward classification tasks, while also maintaining reasonable efficacy for more subtle predictions, like identifying low motivation.

16. Response behaviors

In the following analyses, we check whether values of Cursor Movement Indices (CMIs) these variables vary between students of different gender and socio-economic status (SES) and whether these differences are further moderated by completing the test part in conditions of attentiveness or lowered motivation.

16.1 Gender

Gender differences in school domains (e.g. mathematics) performance in international large-scale assessments (ILSAs) are regularly observed across countries, assessments and over time. While different hypotheses exist about possible sources for these observed differences, empirical evidence is lacking.

Next to sociocultural (e.g. gender stereotypes and norms) and biocognitive hypotheses (e.g. specific cognitive abilities), differences in test-taking behavior are assumed to be a potential source for the observed gender gaps, possibly moderating sociocultural and biological influences (Beller & Gafni, 2000; Bezirhan et al., 2000; DeMars et. al. 2013; Penk, & Richter, 2017). Moreover, there is empirical evidence of gender differences in how students react to manipulations aimed at enhancing their test-related motivation (Finney et al., 2024).

Our study aimed to contribute to the debate on whether female and male examinees in student populations use different problem solving and test-taking strategies or show differences in motivation and test-taking effort which could explain the observed differences in cognitive skills, here, mathematics and science performance.

The data was analysed similarly as in the case of main analysis. Exclusions of students not following instructions were applied in the same way as in other analyses described in this report. Moreover, only students who declared their gender as either male or female were included in this analysis. In the whole sample (before aforementioned exclusions) there were 93 students, i.e. 7.7%, who refused to answer the question about their gender. This reduced the number of analyzed student-test parts[2] in the following analysis from 1956 to 1819 students (i.e. by 7.0%).

[2] Our study had within-subject design, so for each student we get two observations: one describing their behaviour while taking the test with the standard (attentive) instruction and another one describing their behaviour while taking the test with the unattentive instruction. Consequently, in the reported analyses observation is defined as by a combination of specific student and condition (attentive vs. inattentive).

Table 15

Gender Differences in Response Behaviors

Measure	Gender main effect			Attentiveness main effect			Interaction: gender X attentiveness		
	F	p	partial η^2	F	p	partial η^2	F	p	partial η^2
log(dX)	3.17	0.075	0.002	22.65	**0.000**	0.007	7.65	**0.006**	0.002
log(dY)	1.33	0.248	0.001	56.95	**0.000**	0.017	5.62	**0.018**	0.002
sqrt(flipsX)	0.97	0.324	0.001	58.89	**0.000**	0.015	3.37	0.067	0.001
sqrt(flipsY)	2.22	0.137	0.002	49.59	**0.000**	0.013	7.24	**0.007**	0.002
log(vX)	5.02	**0.025**	0.004	143.66	**0.000**	0.048	0.64	0.425	0.000
log(vY)	6.42	**0.011**	0.005	90.04	**0.000**	0.027	0.77	0.381	0.000
log(aX)	12.93	**0.000**	0.010	150.10	**0.000**	0.048	2.77	0.096	0.001
log(aY)	11.96	**0.001**	0.009	97.11	**0.000**	0.029	2.60	0.107	0.001
log(E(response time))	12.47	**0.000**	0.017	211.94	**0.000**	0.086	0.54	0.463	0.000

Notes: N = 1819. Analysis contains both pilot and main study samples. Test statistics and partial η^2 for main effects computed in the absence of the interaction effect (type II test). P-values lower or equal to 0.05 marked in bold.

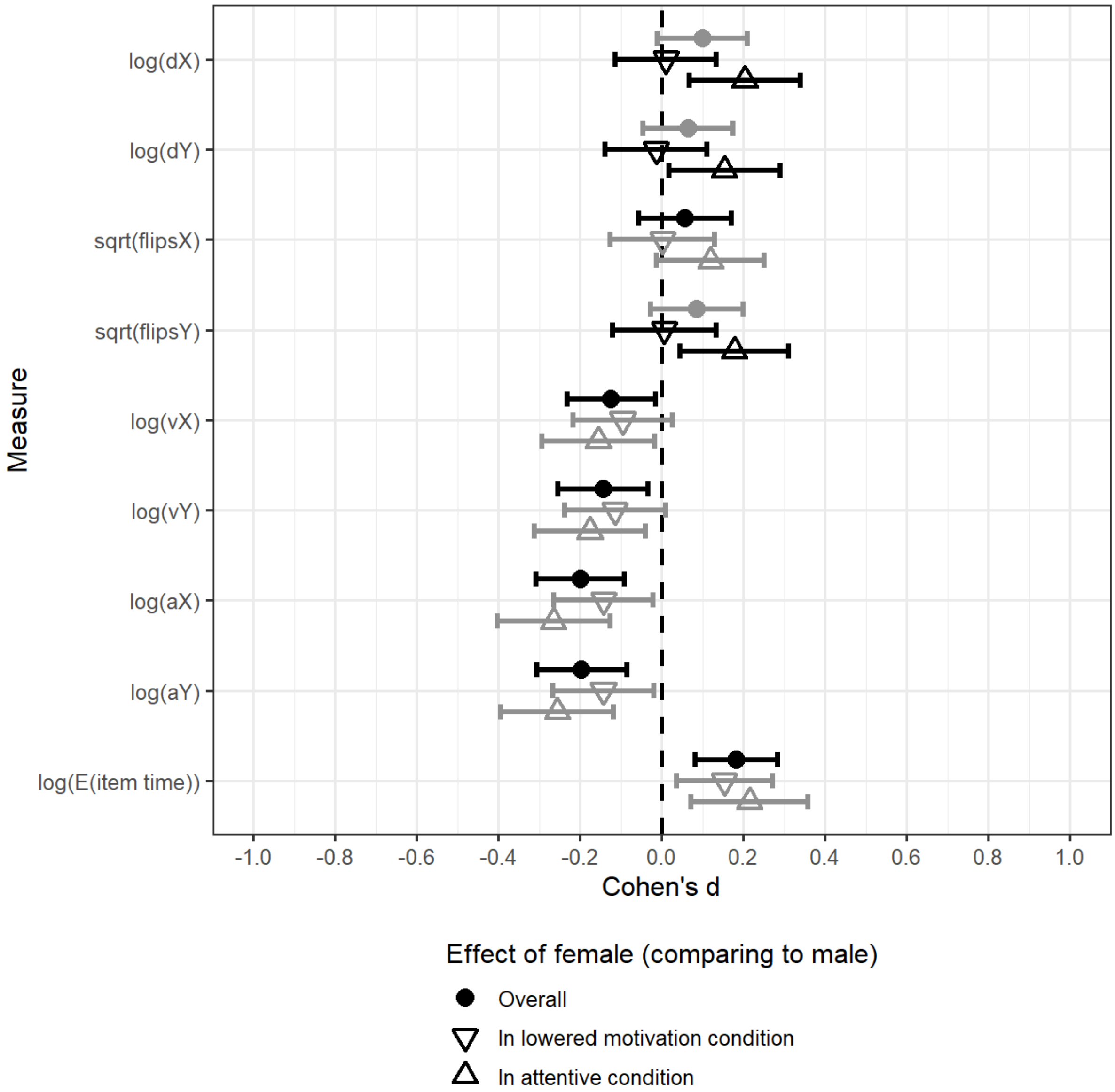
Table 15 presents the results of ANOVAs testing gender differences in response behaviors, focusing on the main effects for gender and attentiveness, as well as their interaction. The results indicate that there are significant differences on several response behaviors between male and female students, including cursor movement speed in both vertical and horizontal dimensions (log(vX), log(vY)), cursor acceleration (log(aX), log(aY)), and screen time (log(E(item time))). As shown in Figure 12, on average, female students showed slower mouse movement speed and lower acceleration. Female participants also spent more time solving test items as indicated by longer screen times.

Interestingly, on average female participants covered more distance with the cursor and had a higher number of vertical flips in the attentive condition. This behavior may suggest that on average they put more effort into responding to items or may use a different strategy for answering, indicating potential variations in approach between genders when fully engaged in the task. This result is to be taken with a certain grain of salt, as the effect size is small and our study may be underpowered to qualify interaction effects.

This pattern seems to point to an effect found previously by some studies that female participants are on average more cautious and risk averse than male respondents in educational tests (e.g. Pekkarinen, 2015; Świst et al., 2015).

Figure 12

Gender differences *(female compared to male) on CMIs and mean item time*



Note: Black bars (as opposed to gray) represent relevant effects – depending on whether interaction of gender and motivation condition was statistically significant at the 0.05 level for a given cursor moves indicator (compare Table 22) these are either separate effects for attentive and lowered motivation condition or a single overall effect. Bars' widths represent 95% confidence intervals. Analysis contains both pilot and main study samples.

16.2 Socio economic status

Studying test-taking motivation among students with varying socio-economic status (SES) is important for both theoretical and practical reasons. Theoretically, it provides insights into how motivational factors interact with SES to influence academic performance and test outcomes. Motivation can mediate or moderate the relationship between SES and performance, offering a more nuanced understanding of educational inequalities. For instance, research on resilient students—those who achieve high academic outcomes despite low SES—highlights the critical role of motivation, self-efficacy, and perseverance (e.g., OECD, 2018). Such studies show that motivation can help offset some of the disadvantages associated with lower SES, making it a key area of interest in understanding resilience mechanisms.

Exclusions of students not following instructions were applied in the same way as in other analyses described in this report. Students' socioeconomic status was operationalized using the question regarding the number of books in a student's household, a measure popularly used as an auxiliary measurement of students' SES in many large-scale low-stakes assessments, including TIMSS. For the purpose of the analysis it was dichotomized to categories "0-25 books" and "25 and more books", which are approximately evenly distributed in the sample (49.3% vs. 50.7% respectively, after exclusions).

Table 16

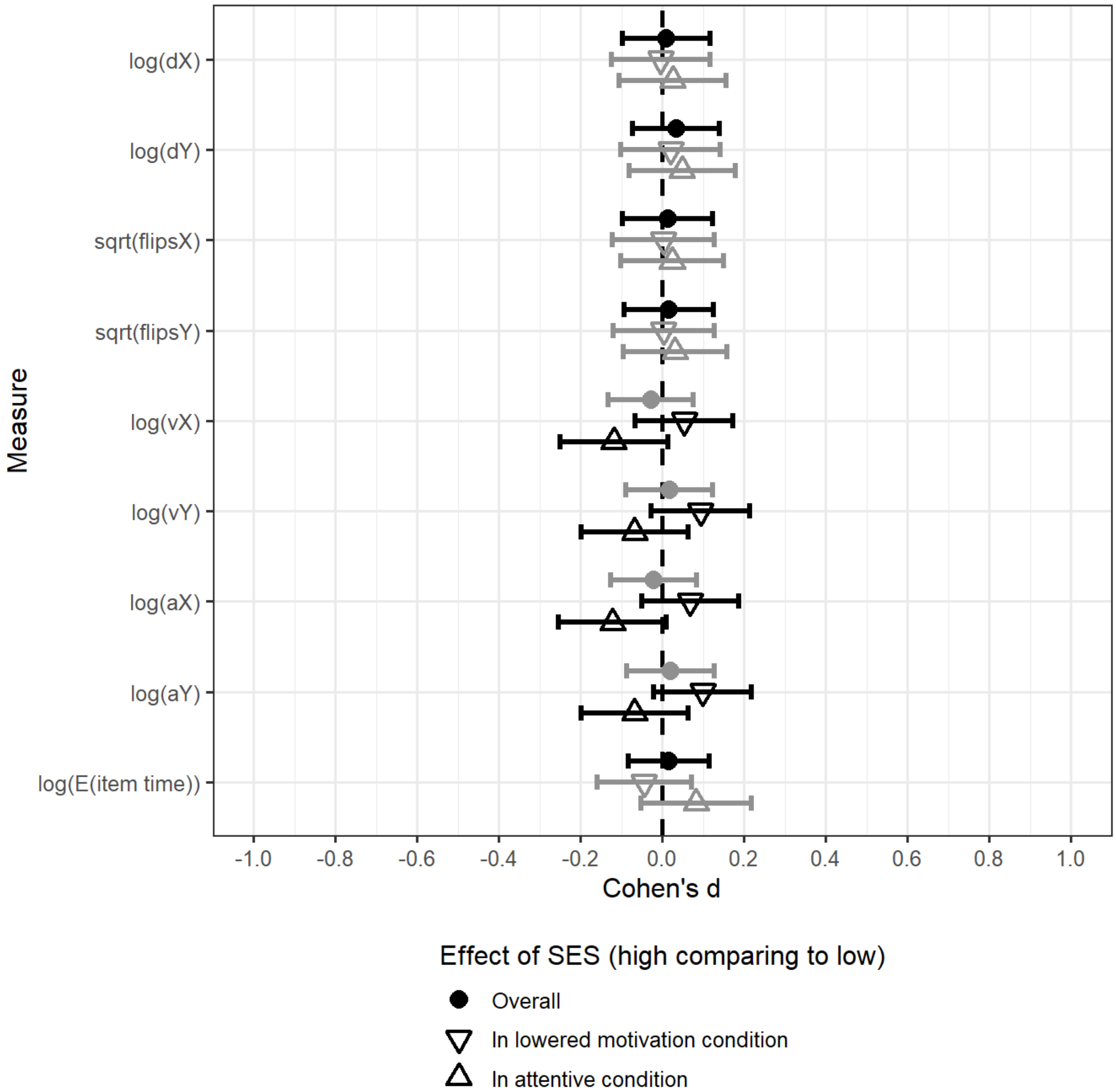
SES Differences in Response Behaviors

Measure	**SES main effect**			**Attentiveness main effect**			**Interaction: SES X attentiveness**		
	F	**p**	**partial η^2**	**F**	**p**	**partial η^2**	**F**	**p**	**partial η^2**
log(dX)	0.04	0.851	0.000	24.47	**0.000**	0.007	0.19	0.660	0.000
log(dY)	0.40	0.528	0.000	60.37	**0.000**	0.017	0.17	0.681	0.000
sqrt(flipsX)	0.06	0.813	0.000	60.39	**0.000**	0.014	0.12	0.732	0.000
sqrt(flipsY)	0.09	0.768	0.000	53.37	**0.000**	0.013	0.22	0.642	0.000
log(vX)	0.27	0.600	0.000	154.67	**0.000**	0.047	5.71	**0.017**	0.002
log(vY)	0.11	0.744	0.000	98.11	**0.000**	0.027	5.72	**0.017**	0.002
log(aX)	0.16	0.687	0.000	161.15	**0.000**	0.047	7.34	**0.007**	0.002
log(aY)	0.14	0.712	0.000	105.92	**0.000**	0.029	6.19	**0.013**	0.002
log(E(response time))	0.11	0.745	0.000	210.83	**0.000**	0.076	2.55	0.110	0.001

Notes: N = 1956. Analysis contains both pilot and main study samples. Test statistics and partial η^2 for main effects computed in the absence of the interaction effect (type II test). P-values lower or equal to 0.05 marked in bold.

Figure 13

Effect of SES on Mouse Movement Indicators and Mean Item Response Time



Note: Black bars (as opposed to gray) represent relevant effects - depending on whether interaction of SES and motivation condition was statistically significant at the 0.05 level for a given indicator (compare Table 16) these are either separate effects for attentive and lowered motivation conditions or a single overall effect. Bars' widths represent 95% confidence intervals. Analysis contains both pilot and main study samples.

No main SES effects on mouse movement indicators were identified. A small interaction effect was identified with high SES participants moving the mouse with less speed and lower acceleration in the attentive condition in contrast to low SES respondents (see Table 16 and Figure 13 for more details).

Research on individual differences in mouse movements and other clickstream data (e.g. Yamauchi et al., 2015; Zhang et al., 2024), as well as on their differences between cognitive or motivational conditions (e.g. Thorpe et al., 2022) is only in its infancy so more studies are needed to understand them better. One of the problems is that many of the studies are conducted in settings that differ greatly from the assessments typically encountered in educational research (Meyer et al., 2024). More studies are needed to gauge individual differences in response processes, as measured by mouse moves and other computer-based paradata. Such differences may stem from socio-demographic (e.g. gender, SES) and cognitive (e.g. level of cognitive ability) participants characteristics and be related to test behavior. Test design is also an important factor that affects response processes and can interact with individual traits of testees (Fernandez-Fontelo et al., 2023; Goldhammer et al., 2017)

17. Discussion

17.1 General Discussion

Our results indicate that, using the Approximate Areas of Interest (AAoIs) approach and a dataset of approximately 1000 students, it is possible to construct a predictive model capable of distinguishing between students completing a cognitive test under lowered versus heightened motivation conditions with an accuracy of approximately 67%.

To put this finding in context, similar methodologies applied to survey research have achieved around 95% accuracy in predicting attentiveness among respondents answering online survey questions. In that setting, however, roughly 75% of respondents were classified as highly attentive, meaning that a naive classifier, assuming all respondents were in a high-attention state, would achieve a baseline accuracy of 75%. Thus, the effective predictive power gain in those studies was around 20% above this baseline, compared to our current study's gain of approximately 15%. This comparison suggests that the differential in predictive strength between these two approaches is, in practical terms, relatively modest.

These values correspond to the values often obtained in the field. For example, Fernandez-Fontelo et al. (2023) used mouse movement data and response time to predict survey variants differing in task complexity, achieving an accuracy of around 60-65% in a number of supervised machine learning models and only around 55% in a neural network model. On the other hand, Schroeders et al. (2022) achieved an accuracy of maximally 70% in their gradient boosting machine model leveraging response patterns and response times and aimed at identifying careless survey respondents. Gogami et al. (2021) achieved an accuracy above 80% in various supervised machine learning models utilizing touchscreen paradata, response patterns, response times, and few types of attention checks, but also reported previous attempts that failed to exceed accuracy of 56%. It seems that still more studies should be conducted to improve modeling capabilities.

Additionally, it is important to consider that the manipulation applied in our study was deliberately subtle, aimed at inducing a moderate decrease in student motivation rather than a complete reduction. Psychometric analysis of test performance reveals only minor differences between the high- and low-motivation conditions, although the test was completed faster and less correctly in the low-motivation condition. This subtle manipulation likely contributes to the challenges in prediction accuracy observed here, as our model deals with discerning less pronounced behavioral indicators of motivation levels.

An interesting aspect of these findings is that, in the absence of explicit knowledge of group assignment (i.e., lowered versus heightened motivation conditions), traditional psychometric methods would not reliably indicate whether students completed the test under lowered motivation or typical conditions. Even under reduced motivation, the test maintains solid psychometric properties, with item parameters remaining relatively stable across conditions. This suggests that, while students may be less engaged, the cognitive assessment itself continues to function robustly, yielding consistent item-level parameters. This stability highlights the subtlety of the motivational manipulation and underscores the necessity of advanced predictive modeling techniques, such as those applied here, to detect nuanced variations in student motivation that standard psychometric indicators alone would likely overlook.

In light of these considerations, the 67% accuracy observed here represents a moderate yet meaningful predictive capability within the constraints of our data and the subtleties of the motivation manipulation. This outcome underscores the complexity of predicting reduced motivation in cognitive testing environments, particularly when differences in engagement levels are relatively nuanced.

17.2 Limitations and Directions for Future Studies

The application of the presented methodology was demonstrated on data where students completed tasks using a computer and mouse, which introduces certain limitations in transferring these methods directly to other testing devices, such as tablets or mobile phones. However, our approach can still be adapted for touchscreen devices, with some adjustments in the data collection process. Leveraging touchscreen data to feed machine learning models aimed at identifying satisficing respondents was demonstrated by Gogami et al. (2021) in the survey context, evidencing that it is feasible.

It seems that the flexibility of our solution is high enough to also leverage data other than mouse moves, e.g. eye-tracking or touchscreen data or even to use data from different sources jointly.

In eye-tracking studies, for instance, data can capture precisely which areas students focus on within the task interface, recording gaze patterns, fixations, and saccades that indicate attention to specific areas of interest (AAOIs). By analyzing these patterns, researchers can gain insights into user behavior similar to those observed with mouse-tracking data, as both capture attention distribution and interaction pathways. Simultaneous tracking of eye-gaze data and mouse moves or touch patterns, would offer a much richer dataset for analysis. Before that will be possible, there is more need for studies relating different data sources with each (i.e. eye-and-mouse tracking studies, e.g. Ding, 2023; Jevermovic et al., 2021).

This adaptability means that, as long as the data collection method allows for tracking attention to specific areas on the screen in real time, the proposed analyses can be effectively conducted with such datasets. Such versatility makes our approach applicable to various assessment formats, including cognitive tests, surveys, and other user-based studies across various digital devices. This capability underscores the robustness of our method, which can support testing and survey applications on diverse devices beyond standard computer-based environments.

The effectiveness and exact effect of motivational instructions is another limitation of the study. First, instructions are mostly aimed at increasing participants' motivation (cf. Huang et al., 2012 or Pokropek et al., 2023 for survey data and Finney et al., 2024 or Rios, 2021 for low-stakes cognitive tests), in contrast to our study in which they were aimed at decreasing it. Of course, as evidenced by our results and other studies using instructions inducing inattentiveness (e.g. Pokropek et al., 2024; Schroeders et al., 2022), such instruction do seem to have an intended effect, however, we cannot be sure whether the obtained results realistically mimic real-world results of low-motivation responding (Ulitzsch et al., 2022).

Using instructions to manipulate participants' motivation also carries risk of respondents who do not comply with them (Ulitzsch et al., 2022). In order to reduce such cases we have employed screening based on manipulation checks and self-reports of test motivation. However, self-reports are inadvertently biased and represent a method of gauging true test effort that is far from perfection. In future, we plan to employ more sophisticated methods of sample screening, based on response patterns and response times (cf. Alfers et al., 2024). Moreover, there is a need to develop intervention methods other than instruction manipulation that would enable to induce the desirable motivational or effort level.

It is also worthy to note a large number of students who failed manipulation checks in this study. This could be a result of overall low motivation to participate in the study, resulting in cursory reading of the instructions or even in adverse reaction to the study and purposeful non-compliance with test procedures. Some students might also feel that the instruction is only a trick and that they should perform attentively anyway. Imperfect measurement properties of the manipulation check itself (only one rating scale item) might also play role in false detections - omissions in case of true non-compliers and false alarms in case of compliers, who failed the manipulation check due to momentary lapse of attention or other factors (e.g. misunderstanding).

Technical details also play an important but largely unknown role in using log-data. Kuric et al. (2024) point out that a variety of study design technical issues can affect mouse tracking data and unknowingly to the researchers warp the results of machine learning models. Literature names features such as mouse speed, sensitivity (represented by CPI - counts per inch), and sampling rate as having an important, yet poorly studied impact on research results (cf. Schoemann et al., 2021). It looks like moving forward the use of mouse-based log-data to study response processes requires empirical studies that would systematically compare various technological settings and indicate their true importance for answering substantial questions.

The design of the study also imposed certain limitations, which limits its generalizability to contexts like ILSAs. ILSAs use a rotated design, which employs a much larger number of items than it was used in this study.

Furthermore, they also utilize diverse item formats beyond multiple-choice, it seems that especially constructed response items can be more influenced by motivational levels, but also allow for an even richer paradata to be gathered (e.g. Ivanova et al., 2020).

Moreover, the measurement instrument was short, shorter than in the case of TIMSS Grade 8 or Polish Grade 8 National Examination. The brevity of the test might also negatively contribute to the possibility of distinguishing between motivated and non-motivated participants and to the feasibility of identifying unique response behaviors.

It is also hard to tell what was the initial level of test-related motivation in the tested sample and how it was comparable to motivation in real ILSAs performed in Poland. A possibility that the level was low initially and the manipulated instruction did not have space to lower it further is somewhat probable. Measuring test motivation at least twice - before and after the test - could help in gathering more information on students' motivation and test perception (Myers & Finney, 2021)

This may point to more applicability of studies that intend to increase instead to decrease the level of students' motivation.

ILSAs are also administered across cultures with unique test-taking and evaluation traditions. This may bring differences not only in motivation levels but also in ways students display high or low motivation in response behavior. Future research should validate the approach across different contexts to enhance generalizability to diverse assessment environments.

17.3 Conclusions

The results presented in this report provide a strong foundation for drafting a research article that will highlight the key findings from our analyses. The article will delve not only into the specifics of the applied approach but also into the practical implications of the findings, especially concerning the potential for predicting and detecting reduced student motivation in cognitive assessments.

Furthermore, the dataset itself appears to be an exceptionally valuable tool for testing solutions aimed at detecting response patterns related to low test-taking motivation. Pending upon IEA's consent, we plan to make this dataset publicly available, enabling researchers to test their methods on a rich experimental dataset containing diverse indicators and a robust range of variables, granting open access to high-quality experimental data for the research community. Hopefully, this will also foster the development of tools for analyzing response processes under various motivations in educational contexts.

Additionally, we intend to reanalyze these data using novel machine learning techniques to verify whether we can achieve even more accurate predictions of reduced student motivation. Our aim is also to expand this analysis by incorporating demographic variables, which could contribute to a deeper understanding of the relationship between motivation and other student characteristics. This study offers extensive applications and serves as a foundation for future analyses that may advance methodologies in large-scale educational assessments.

Further exploration of these data with cutting-edge machine learning solutions will allow us to evaluate the effectiveness of new algorithms in identifying subtle patterns of decreased engagement. As technology progresses, these types of analyses could provide innovative insights into motivational dynamics, ultimately supporting educators, policymakers, and researchers in creating more adaptive and responsive assessment environments.

Acknowledgments and Advisory Contributions

The details of the study's design were presented to a group of external experts before the study's implementation. The group of experts included Esther Ulitzsch, Ulf Kröhne, and Francesca Borgonovi. The project's presentation and design were initially shared during online meetings, followed by ongoing consultations via email. All decisions related to the study's design, planning, and technical details concerning the collected data were developed by the research team and consulted with the experts, who had only an advisory role. The research team is solely responsible for any errors and imperfections in the study.

We would like to acknowledge and sincerely thank Esther Ulitzsch, Ulf Kröhne, and Francesca Borgonovi for their valuable time and insightful advice during the consultation process. Their input was instrumental in refining the study's approach and ensuring a thoughtful and informed design.

References

Alfers, T., Gittler, G., Ulitzsch, E., & Pohl, S. (2024). Assessing the Speed–Accuracy Tradeoff in Psychological Testing Using Experimental Manipulations. *Educational and Psychological Measurement*, 00131644241271309.

Arslan, B., & Finn, B. (2023). The effects of personalized nudges on cognitively disengaged student behavior in low-stakes assessments. *Journal of Intelligence, 11*(11), 204.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? Sex Roles, 42(1–2), 1–21. http://doi.org/10.1023/A:1007051109754

Bezirhan, U., von Davier, M. & Grabovsky, I. (2020). Modeling Item Revisit Behavior: The Hierarchical Speed–Accuracy–Revisits Model. Educational and Psychological Measurement, X, 1-25. https://doi.org/10.1177/0013164420950556

Brenner, E., & Smeets, J. B. (2003). Fast corrections of movements with a computer mouse. Spatial Vision, 16(3), 365–376

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Clark, J. W., & Stephane, A. L. (2018). Affordable Eye Tracking for Informed Web Design. In A. Marcus & W. Wang (Eds.), Design, User Experience, and Usability: Theory and Practice (pp. 346–355). Springer International Publishing.

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. Research & Practice in Assessment, 8, 69–82. https://eric.ed.gov/?id=EJ1062839

Ding, C. (2023). *Mouse Tracking for Reading (MoTR): A New Incremental Processing Paradigm* (Master's thesis, University of Zurich). https://www.zora.uzh.ch/id/eprint/259944/

Fernández-Fontelo, A., Kieslich, P. J., Henninger, F., Kreuter, F., & Greven, S. (2023). Predicting question difficulty in web surveys: A machine learning approach based on mouse movement features. *Social Science Computer Review*, *41*(1), 141-162.

Finney, S. J., Schaefer, K. E., & McFadden, M. E. (2024). Priming Examinees to Give Good Effort: Differential Utility Across Gender Identity. *The Journal of Experimental Education*, 1-18.

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database* (2nd ed.). TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/timss2019/international-database/

Gogami, M., Matsuda, Y., Arakawa, Y., & Yasumoto, K. (2021). Detection of careless responses in online surveys using answering behavior on smartphone. *IEEE Access*, *9*, 53205-53218.

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments.* https://doi.org/10.1186/s40536-017-0051-9

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks.* Springer.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horwitz, R., Brockhaus, S., Henninger, F., Kieslich, P. J., Schierholz, M., Keusch, F., & Kreuter, F. (2020). Learning from mouse movements: Improving questionnaires and respondents' user experience through passive data collection. *Advances in Questionnaire Design, Development, Evaluation and Testing,* 403-425.

Horwitz, R., Kreuter, F., & Conrad, F. (2017). Using mouse movements to predict web survey response difficulty. *Social Science Computer Review, 35*(3), 388-405.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27,* 99-114.

IEA (2019). *TIMSS 2019 international databases.* Retrieved from https://timss2019.org/international-database/

Ivanova, M., Michaelides, M., & Eklöf, H. (2020). How does the number of actions on constructed-response items relate to test-taking effort and performance?. Educational Research and Evaluation, 26(5-6), 252-274.

Jevremovic, A., Zaphiris, P., Adamovic, S., Mottus, M., & Ioannou, A. (2021). Applying information theory and entropy to eliminate errors in mouse-tracking results. *arXiv preprint arXiv:2105.06320*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations.*

Kirsh, I., & Joy, M. (2020). Exploring Pointer Assisted Reading (PAR): Using Mouse Movements to Analyze Web Users' Reading Behaviors and Patterns. In C.

Stephanidis, M. Kurosu, H. Degen, & L. Reinerman-Jones (Eds.), HCI International 2020—Late Breaking Papers: Multimodality and Intelligence (pp. 156–173). Springer International Publishing.

Kuric, E., Demcak, P., Krajcovic, M., & Nemcek, P. (2024). Is mouse dynamics information credible for user behavior research? An empirical investigation. *Computer Standards & Interfaces,* 90, 103849.

Leipold, F. M., Kieslich, P. J., Henninger, F., Fernández-Fontelo, A., Greven, S., & Kreuter, F. (2024). Detecting respondent burden in online surveys: How different sources of question difficulty influence cursor movements. *Social Science Computer Review*, 08944393241247425.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment, 20*(2), 79–94. https://doi.org/10.1080/10627197.2015.1028618

Meidenbauer, K. L., Niu, T., Choe, K. W., Stier, A. J., & Berman, M. G. (2023). Mouse movements reflect personality traits and task attentiveness in online experiments. *Journal of Personality, 91*(2), 413-425.

Michaelides, M. P., & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: Country and item-type differences. *Psychological Test and Assessment Modeling, 64*(3), 304-338.

Myers, A. J., & Finney, S. J. (2021). Change in self-reported motivation before to after test completion: Relation with performance. The Journal of Experimental Education, 89(1), 74-94.

OECD (2018). Equity in Education: Breaking Down Barriers to Social Mobility. OECD Publishing. https://doi.org/10.1787/9789264073234-en

Pekkarinen, T. (2015). Gender differences in behavior under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization, 115,* 94-110.

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. Educational Assessment, Evaluation and Accountability, 29(1), 55-79. https://doi.org/10.1007/s11092-016-9248-7

Pokropek, A., Żółtak, T., & Muszyński, M. (2023). Mouse chase: Detecting careless and unmotivated responders using cursor movements in web-based surveys. *European Journal of Psychological Assessment, 39*(4), 299-306.

Pokropek, A., Żółtak, T., & Muszyński, M. (2024). Identifying careless responding in web-based surveys: Exploiting sequence data from cursor trajectories and approximate areas of interest. *Zeitschrift für Psychologie, 232*(2), 95.

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education, 34*(2), 85-106.

Schoemann, M., O'Hora, D., Dale, R., & Scherbaum, S. (2021). Using mouse cursor tracking to investigate online cognition: Preserving methodological ingenuity while moving toward reproducible science. *Psychonomic Bulletin & Review*, *28*(3), 766-787.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems* (pp. 2951–2959).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958.

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. Computers in Human Behavior, 26(6), 1488–1495.

Świst, K., Skórska, P., Koniewski, M., & Jasińska-Maciążek, A. (2015). Sex differences in guessing and item omission. *Edukacja, 3*(134), 48-62.

Thorpe, A., Friedman, J., Evans, S., Nesbitt, K., & Eidels, A. (2021). Mouse movement trajectories as an indicator of cognitive workload. *International Journal of Human–Computer Interaction, 38*(15), 1464–1479. https://doi.org/10.1080/10447318.2021.2002054

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593-619.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology, 67*(2), 231-263.

Ward, M. K., & Pond III, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior, 48,* 554-568.

Yamauchi, T., Seo, J. H., Jett, N., Parks, G., & Bowman, C. (2015). Gender differences in mouse and cursor movements. *International Journal of Human–Computer Interaction, 31*(12), 911–921.

Yamauchi, T., & Xiao, K. (2018). Reading emotion from mouse cursor motions: Affective computing approach. Cognitive Science, 42(3), 771–819.

Yen, W. M. (1982). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Zhang, S., Wilmut, K., Zhang, K., & Wang, S. (2024). Age-related changes in motor planning for prior intentions: A mouse tracking reach-to-click task. *Frontiers in Psychology, 15,* 1323798.

Żółtak, T. (2024). *tzoltak/logLime: Version 0.3.1* (v0.3.1). Zenodo. https://doi.org/10.5281/zenodo.12087493

Annex

Annex 1

Instructions and Questionnaire – Original Form

Instrukcja dla Części Testu z Zadaniami do Rozwiązania Poprawnie – wersja kolejności Attentive-Inattentive

Cześć!

Zaczynamy nasz test od bardzo ważnej części. To szansa, żebyś mógł/mogła sprawdzić swoje umiejętności i przygotować się do przyszłych egzaminów.

Co musisz zrobić?

Prosimy Cię, abyś teraz skupił/skupiła się na zadaniach i rozwiązał/rozwiązała je najlepiej, jak potrafisz. Podejdź do tego testu poważnie, traktując go jako próbę przed prawdziwymi wyzwaniami.

Dlaczego to ważne?

Twoje wyniki z tej części testu będą dla Ciebie dostępne. To świetna okazja, żeby zobaczyć, gdzie jesteś mocny/mocna i nad czym warto jeszcze popracować.

O czym pamiętać?

- **Staraj się:** Twoje zaangażowanie jest kluczowe. Pokaż, na co Cię stać!
- **Obliczenia w głowie:** Prosimy, abyś próbował/próbowała robić wszystkie obliczenia samodzielnie, bez używania kartek, kalkulatorów czy innych pomocy.
- **Samodzielność:** To Twoja szansa, aby zobaczyć, jak radzisz sobie bez zewnętrznej pomocy.

To dla nas ważne, abyś dał/dała z siebie wszystko. Jesteśmy przekonani, że poradzisz sobie świetnie i będziesz mógł/mogła być dumny/dumna z własnych osiągnięć. Na samym końcu badania otrzymasz informację o wyniku, jaki uzyskałeś/uzyskałaś.

Gotowy/gotowa do startu? Pamiętaj, że każde zadanie to nowa możliwość nauki. Powodzenia!

Instrukcja dla Części Testu z Zadaniami do Rozwiązania Poprawnie – wersja kolejności Attentive-Inattentive

Witaj ponownie!

Teraz przechodzimy do bardzo ważnej części naszego testu. To szansa, żebyś mógł/mogła sprawdzić swoje umiejętności i przygotować się do przyszłych egzaminów.

Co musisz zrobić?

Prosimy Cię, abyś teraz skupił/skupiła się na zadaniach i rozwiązał/rozwiązała je najlepiej, jak potrafisz. Podejdź do tego testu poważnie, traktując go jako próbę przed prawdziwymi wyzwaniami.

Dlaczego to ważne?

Twoje wyniki z tej części testu będą dla Ciebie dostępne. To świetna okazja, żeby zobaczyć, gdzie jesteś mocny/mocna i nad czym warto jeszcze popracować.

O czym pamiętać?

- **Staraj się:** Twoje zaangażowanie jest kluczowe. Pokaż, na co Cię stać!
- **Obliczenia w głowie**: Prosimy, abyś próbował/próbowała robić wszystkie obliczenia samodzielnie, bez używania kartek, kalkulatorów czy innych pomocy.
- **Samodzielność**: To Twoja szansa, aby zobaczyć, jak radzisz sobie bez zewnętrznej pomocy.

To dla nas ważne, abyś dał/dała z siebie wszystko. Jesteśmy przekonani, że poradzisz sobie świetnie i będziesz mógł/mogła być dumny/dumna z własnych osiągnięć.

Na samym końcu badania otrzymasz informację o wyniku, jaki uzyskałeś/uzyskałaś.

Gotowy/gotowa do startu? Pamiętaj, że każde zadanie to nowa możliwość nauki. Powodzenia!

Instrukcja do nieuważnego rozwiązywania – wersja kolejności Attentive-Inattentive

Witaj ponownie!

Teraz mamy dla Ciebie inne zadanie. Te pytania są specjalne i chcemy, żebyś podszedł/podeszła do nich w inny sposób.

O co chodzi?

Chcemy, żebyś spróbował/spróbowała odpowiedzieć na te pytania, nie wkładając w to zbyt wiele wysiłku. Wyobraź sobie, że nie musisz się za bardzo starać i że wynik tych zadań w ogóle się nie liczy. Tak jakbyś miał/miała zrobić coś na szybko, bez większego zastanowienia.

Dlaczego to robimy?

To część naszego badania. Chcemy zobaczyć, jak ludzie odpowiadają na pytania, kiedy nie mają motywacji, aby się starać. **Nie martw się, te odpowiedzi nie wpłyną na Twój wynik w teście.**

Pamiętaj:

- Nie musisz się starać.
- Nie zastanawiaj się zbyt długo nad każdym pytaniem.
- Twoje odpowiedzi tutaj nie będą oceniane.

Gotowy/Gotowa? Zaczynajmy i pamiętaj, że to tylko część testu. Powodzenia!

Instrukcja do nieuważnego rozwiązywania – wersja kolejności Inattentive-Attentive

Cześć!

Mamy dla Ciebie specjalne zadanie. Te pytania są inne niż wszystkie i chcemy, żebyś podszedł/podeszła do nich w nietypowy sposób.

O co chodzi?

Chcemy, żebyś spróbował/spróbowała odpowiedzieć na te pytania, nie wkładając w to zbyt wiele wysiłku. Wyobraź sobie, że nie musisz się za bardzo starać i że wynik tych zadań w ogóle się nie liczy. Tak jakbyś miał/miała zrobić coś na szybko, bez większego zastanowienia.

Dlaczego to robimy?

To część naszego badania. Chcemy zobaczyć, jak ludzie odpowiadają na pytania, kiedy nie mają motywacji, aby się starać. Nie martw się, te odpowiedzi nie wpłyną na Twój wynik w teście.

Pamiętaj:

- Nie musisz się starać.
- Nie zastanawiaj się zbyt długo nad każdym pytaniem.
- Twoje odpowiedzi tutaj nie będą oceniane.

Gotowy/Gotowa? Zaczynajmy i pamiętaj, że to tylko część testu. Powodzenia!

Instructions and Questionnaire – Translated to English

Instructions for the Test Section with Correctly Solving Tasks – Order Version Attentive-Inattentive

Hello!

We are starting our test with a very important part. This is your chance to check your skills and prepare for future exams.

What do you need to do?

We ask you to focus on the tasks now and solve them as best as you can. Approach this test seriously, treating it as a practice for real challenges.

Why is this important?

Your results from this part of the test will be available to you. This is a great opportunity to see where your strengths lie and what areas you might need to work on.

What to remember?

- Engagement: Your engagement is crucial. Show what you are capable of!
- Mental Calculations: Please try to do all calculations on your own, without using paper, calculators, or other aids.
- Independence: This is your chance to see how well you can handle things without external help.

It is important to us that you give it your all. We are confident that you will do great and be proud of your achievements. At the end of the study, you will receive information about your score.

Ready to start? Remember, each task is a new learning opportunity. Good luck!

Instructions for the Test Section with Correctly Solving Tasks – Order Version Attentive-Inattentive

Hello again!

Now we move on to a very important part of our test. This is your chance to check your skills and prepare for future exams.

What do you need to do?

We ask you to focus on the tasks now and solve them as best as you can. Approach this test seriously, treating it as a practice for real challenges.

Why is this important?

Your results from this part of the test will be available to you. This is a great opportunity to see where your strengths lie and what areas you might need to work on.

What to remember?

- Engagement: Your engagement is crucial. Show what you are capable of!
- Mental Calculations: Please try to do all calculations on your own, without using paper, calculators, or other aids.
- Independence: This is your chance to see how well you can handle things without external help.

It is important to us that you give it your all. We are confident that you will do great and be proud of your achievements. At the end of the study, you will receive information about your score.

Ready to start? Remember, each task is a new learning opportunity. Good luck!

Instructions for Inattentive Solving – Order Version Attentive-Inattentive

Hello again!

Now we have a different task for you. These questions are special, and we want you to approach them differently.

What is it about?

We want you to try answering these questions without putting in much effort. Imagine that you do not need to try too hard and that the outcome of these tasks does not matter at all. Just like doing something quickly without much thought.

Why are we doing this?

This is part of our study. We want to see how people respond to questions when they are not motivated to try hard. Do not worry; these answers will not affect your test score.

Remember:

- You do not need to try hard.
- Do not think too long about each question.
- Your answers here will not be graded.

Ready? Let's start and remember, this is just part of the test. Good luck!

Instructions for Inattentive Solving – Order Version Inattentive-Attentive

Hello!

We have a special task for you. These questions are different from all the others, and we want you to approach them in an unusual way.

What is it about?

We want you to try answering these questions without putting in much effort. Imagine that you do not need to try too hard and that the outcome of these tasks does not matter at all. Just like doing something quickly without much thought.

Why are we doing this?

This is part of our study. We want to see how people respond to questions when they are not motivated to try hard. Do not worry; these answers will not affect your test score.

Remember:

- You do not need to try hard.
- Do not think too long about each question.
- Your answers here will not be graded.

Ready? Let's start and remember, this is just part of the test. Good luck!

Manipulation check – after Inattentive Instruction

Przed rozpoczęciem, chcemy się upewnić, czy zrozumiałeś/aś instrukcję. Jak zamierzasz podejść do tych zadań?

a) Mam odpowiedzieć na pytania, nie wkładając w to zbyt wiele wysiłku, mogę zgadywać.

b) Mam rozwiązywać zadania najlepiej, jak potrafię.

c) Mam je przeczytać, ale nie jestem pewien/pewna, co mam zrobić.

d) Mam ignorować instrukcję i działać według własnego uznania.

Translated:

Before we begin, we want to ensure that you understand the instructions. How do you plan to approach these tasks?

A. I should answer the questions without putting in much effort; I can guess.

B. I should solve the tasks as best as I can.

C. I should read them, but I am not sure what to do.

D. I should ignore the instructions and act according to my own judgment.

Questionnaire (Kwestionariusz) - Original Form

Prosimy o dokładne wypełnienie poniższego kwestionariusza. Twoje odpowiedzi są bardzo ważne i pozwolą nam ulepszyć przyszłe testy. Wszystkie informacje są zbierane anonimowo. Wyniki badania zostaną wykorzystane tylko w tekstach naukowych.

1. Do której chodzisz klasy:

a. 7

b. 8

2. Płeć:

a. Chłopak

b. Dziewczyna

c. Wolę nie odpowiadać

3. Jak bardzo starałeś/starałaś się dać z siebie wszystko, aby poprawnie wypełnić test, w którym prosiliśmy cię o poprawne odpowiadania na zadania?

a. Bardzo się starałem/starałam

b. Raczej się starałem/starałam

c. Mało się starałem/starałam

d. W ogóle się nie starałem/starałam

4. Jaka była Twoja strategia odpowiadania na pytania, w których prosiliśmy cię abyś zgadywał/zgadywała?

a. Po prostu klikałem/klikałam jak najszybciej pierwszą lepszą odpowiedź.

b. Zaznaczałem/zaznaczałam wszędzie tę samą odpowiedź.

c. Czytałem pobieżnie i starałem się wyeliminować najmniej poprawne odpowiedzi.

d. Dokładnie czytałem pytania i starannie analizowałem wszystkie dostępne odpowiedzi przed wyborem.

e. Odpowiadałem w inny sposób [otwarte]

5. Czy używałeś/aś brudnopisu podczas testu?

a. Tak

b. Nie

6. Czy używałeś/aś kalkulatora podczas testu?

a. Tak

b. Nie

7. Czy korzystałeś/aś z telefonu podczas testu (np. do obliczeń lub wyszukiwania informacji)?

a. Tak

b. Nie

8. Czy korzystałeś/aś z przeglądarki internetowej na komputerze podczas testu (np. do obliczeń lub wyszukiwania informacji)?

a. Tak

b. Nie

9. Czy napotkałeś/aś jakiekolwiek problemy techniczne? Zaznacz wszystkie poprawne odpowiedzi?

a. Nie, wszystko było w porządku

b. Test się zacinał

c. Test był niewyraźny, trudno było odczytać niektóre opisy lub rysunki

d. Myszka źle działała

10. Na ile dokładnie ankieter przedstawił Ci zasady dotyczące tego badania?

a. Bardzo dokładnie

b. Raczej dokładnie

c. Raczej niedokładnie

d. Bardzo niedokładnie

e. W ogóle ich nie przedstawił

11. Ile mniej więcej książek znajduje się w Twoim domu? (Nie licząc czasopism, gazet ani książek szkolnych. Zaznacz tylko jedno pole.

a. Kilka książek lub wcale ich nie ma (0–10 książek)

b. Wystarczająco, aby wypełnić jedną półkę (11–25 książek)

c. Wystarczająco, aby wypełnić jedną biblioteczkę (26–100 książek

d. Wystarczająco, aby wypełnić dwie biblioteczki (101–200 książek)

e. Wystarczająco, aby wypełnić trzy lub więcej biblioteczek (ponad 200 książek)

Questionnaire - Translated Form

Please carefully fill out the questionnaire below. Your answers are very important and will help us improve future tests. All information is collected anonymously. The results of the study will be used only in academic texts.

1. **What grade are you in?**
    a. 7
    b. 8
2. **Gender:**
    a. Boy
    b. Girl
    c. Prefer not to say
3. **How much effort did you put into trying to correctly complete the test where we asked you to answer the tasks correctly?**
    a. I tried very hard
    b. I tried fairly hard
    c. I put in little effort
    d. I did not try at all
4. **What was your strategy for answering the questions where we asked you to guess?**
    a. I just clicked the first answer as quickly as possible.
    b. I marked the same answer everywhere.
    c. I skimmed the questions and tried to eliminate the least correct answers.
    d. I read the questions carefully and thoroughly analyzed all available answers before choosing.
    e. I answered in another way [open-ended]
5. **Did you use scratch paper during the test?**
    a. Yes

b. No

6. **Did you use a calculator during the test?**

   a. Yes

   b. No

7. **Did you use your phone during the test (e.g., for calculations or looking up information)?**

   a. Yes

   b. No

8. **Did you use a web browser on your computer during the test (e.g., for calculations or looking up information)?**

   a. Yes

   b. No

9. **Did you encounter any technical problems? Check all that apply:**

   a. No, everything was fine

   b. The test was lagging

   c. The test was blurry, it was hard to read some descriptions or drawings

   d. The mouse was malfunctioning

10. **How clearly did the interviewer explain the rules of this study to you?**

    a. Very clearly

    b. Fairly clearly

    c. Fairly unclearly

    d. Very unclearly

    e. Did not explain at all

11. **Approximately how many books are there in your home? (Do not count magazines, newspapers, or schoolbooks. Check only one box.)**

    a. A few books or none at all (0–10 books)

    b. Enough to fill one shelf (11–25 books)

c. Enough to fill one bookcase (26–100 books)

d. Enough to fill two bookcases (101–200 books)

e. Enough to fill three or more bookcases (over 200 books)

Information at the end of the study – Original Form

To już koniec testu!

Dziękujemy Ci bardzo za poświęcenie czasu i wypełnienie naszego testu i kwestionariusza. Chcemy Cię poinformować, że ostatnie pytanie testowe w części, w której prosiliśmy cię o poprawne wypełnianie testu było specjalnie skonstruowane tak, aby zawierało trzy poprawne odpowiedzi.

To pytanie było częścią naszego badania i miało na celu sprawdzenie reakcji na nietypową sytuację w teście. Odpowiedź na to pytanie nie będzie miała wpływu na Twój końcowy wynik.

W teście zbieraliśmy również informacje o czasie odpowiedzi i ruchach myszką na ekranie. Są to dane, które pomogą nam projektować lepsze testy w przyszłości.

Twój wynik na teście to XXXX% poprawnych odpowiedzi

Jesteśmy wdzięczni za Twoje uczestnictwo i zaangażowanie w procesie badawczym. Twoje odpowiedzi są dla nas bardzo cenne i pomogą w lepszym zrozumieniu, jak uczniowie podchodzą do różnych typów pytań testowych.

Jeszcze raz dziękujemy i życzymy Ci wszelkiej pomyślności w dalszej edukacji!

Information After Completing the Study – Translated

This is the end of the test!

Thank you very much for taking the time to complete our test and questionnaire. We would like to inform you that the last test question in the section where we asked you to answer correctly was specifically designed to contain three correct answers.

This question was part of our study and aimed to observe reactions to an unusual situation in the test. Your answer to this question will not affect your final score.

During the test, we also collected data on response times and mouse movements on the screen. This data will help us design better tests in the future.

Your test score is XXXX% correct answers.

We are grateful for your participation and engagement in the research process. Your responses are very valuable to us and will help in better understanding how students approach different types of test questions.

Once again, thank you, and we wish you all the best in your future education!

Annex 2

Selection of items for the math test

Table A2.1 displays key psychometric characteristics of the items selected for the math test in the current study. TIMSS items characteristics come from the TIMSS documentation (Fishbein et al., 2019; IEA, 2019), while Polish grade 8 exam item characteristics were calculated by the project's PI on the basis of publicly available data applying the 3PLM IRT model. Although the two tests were not linked, the displayed item characteristics convey the message about their general difficulty and other properties.

Table A2.1

Psychometric Parameters of the Selected Items

Item source	Content domain	Original number	*a*	*b*	*c*
TIMSS	Number	M042032	0.86	-0.51	0.24
		M042041	1.40	-0.22	0.33
		M032166	1.07	0.02	0.19
		M032094	1.33	0.06	0.29
		M042031	1.62	0.52	0.24
		M032704	1.21	-0.09	0.16
	Data & Chance	M042254	0.82	-0.97	0.22
		M022101	1.31	0.01	0.20
		M042260	1.03	0.04	0.30
		M042222	1.04	0.49	0.10
	Algebra	M052302	0.97	-0.70	0.10
		M032295	1.37	-0.50	0.26
		M032738	1.36	-0.19	0.24
		M042236	1.47	0.24	0.25
		M032477	1.83	0.59	0.24
		M032198	1.41	0.64	0.25
	Geometry	M032579	1.17	0.00	0.20
		M052084	1.46	0.27	0.13

Item source	Content domain	Original number	*a*	*b*	*c*
		M032679	1.22	0.29	0.23
		M032116	1.23	0.78	0.26
PL Exam	Number	gm_m_2016_z03	1.64	-0.50	0.19
		gm_m_2018_z07	2.36	-0.06	0.27
		gm_m_2014_z11	2.33	-0.35	0.22
	Data & Chance	gm_2002_z05	1.74	-1.31	0.18
		gm_m_2016_z01	1.86	-0.76	0.22
	Algebra	e8_2022_z7	1.40	-0.17	0.21
		gm_m_2015_z07	0.87	-0.89	0.06
		gm_2006_z08	1.63	1.66	0.05
	Geometry	gm_m_2016_z20	1.53	-0.83	0.21
		gm_2004_z04	1.66	-0.26	0.24

The allocation of items to different versions of the test is presented in Table A2.2. The items were selected to ensure that each part of the test has a similar level of difficulty and an even distribution of content domains to make the two test versions as comparable as possible.

Table A2.2

Item Allocation to Test Versions

	Version A	Version B
Part	**Item**	**Item**
1	M032579	M042032
1	M042254	M052302
1	M042222	gm_m_2016_z20
1	gm_2006_z08	M032738
1	M042260	M032477
1	gm_m_2015_z07	gm_m_2018_z07
1	M032295	M032704
1	M042031	M032116
1	gm_2004_z04	gm_m_2016_z01
1	M032679	M042236
1	M032198	gm_2002_z05

	Version A	Version B
Part	**Item**	**Item**
1	gm_m_2014_z11	gm_m_2016_z03
1	M042041	M022101
1	e8_2022_z7	M052084
1	M032166	M032094
New instructions		
2	M042032	M032579
2	M052302	M042254
2	gm_m_2016_z20	M042222
2	M032738	gm_2006_z08
2	M032477	M042260
2	gm_m_2018_z07	gm_m_2015_z07
2	M032704	M032295
2	M032116	M042031
2	gm_m_2016_z01	gm_2004_z04
2	M042236	M032679
2	gm_2002_z05	M032198
2	gm_m_2016_z03	gm_m_2014_z11
2	M022101	M042041
2	M052084	e8_2022_z7
2	M032094	M032166

Annex 3

Pilot administration data collection process

Table A3.1

Overview of the Pilot administration

Day	Number of sessions	Overall number of students	Number of accepted records	Cumulative number of accepted records	Number of excluded records					
					Overall	NC	S	LBT	MM	T
2024-04-08	3	39	32	32	7	0	2	1	1	3
2024-04-09	2	16	13	45	3	0	3	0	0	0
2024-04-12	2	23	21	66	2	1	0	0	1	0
Overall	7	78	66	66	12	1	5	1	2	3

Notes:
Number of accepted records = Overall number of students – Overall number of excluded records
Reasons for excluding records:
NC – not reaching the end of the survey,
S – completing the survey in time below 10 minutes (speeding),
LBT – leaving the browser tab with the survey during the survey completion,
MM – the same login (token) was used on different computers,
T —analysis of responses and timing of responses indicates that it is highly probable that two or more students took the test together

The criterion used to determine that a pair of students probably took the test together was fulfilling all of the following conditions:

- The difference between test start times less than 60 seconds,
- The sum (across all the survey screens) of the absolute values of differences between cumulative screen times below 15 seconds.
- At least 13 (out of 15) responses to the items in the item set with attentive instruction were the same.
- At least 13 (out of 15) responses to the items in the item set with inattentive instruction were the same.

Table A3.2

Demographic characteristics of the Pilot administration sample

	Overall		Attentive first, set one first		Attentive first, set two first		Inattentive first, set one first		Inattentive first, set two first	
	N	Pct.	N	Pct.	N	Pct.	N	Pct.	N	Pct.
Gender										
Don't want to answer	6	9.1%	0	0.0%	1	6.3%	2	22.2%	3	12.0%
Female	22	33.3%	5	31.3%	5	31.3%	3	33.3%	9	36.0%
Male	38	57.6%	11	68.8%	10	62.5%	4	44.4%	13	52.0%
Locality size										
Village	66	100.0%	16	100.0%	16	100.0%	9	100.0%	25	100.0%
Grade										
7	39	59.1%	13	81.3%	0	0.0%	9	100.0%	17	68.0%
8	27	40.9%	3	18.8%	16	100.0%	0	0.0%	8	32.0%
Overall	66	100.0%	16	100.0%	16	100.0%	9	100.0%	25	100.0%

Main study data collection process

Table A3.1

Data Collection Flow in Main Study

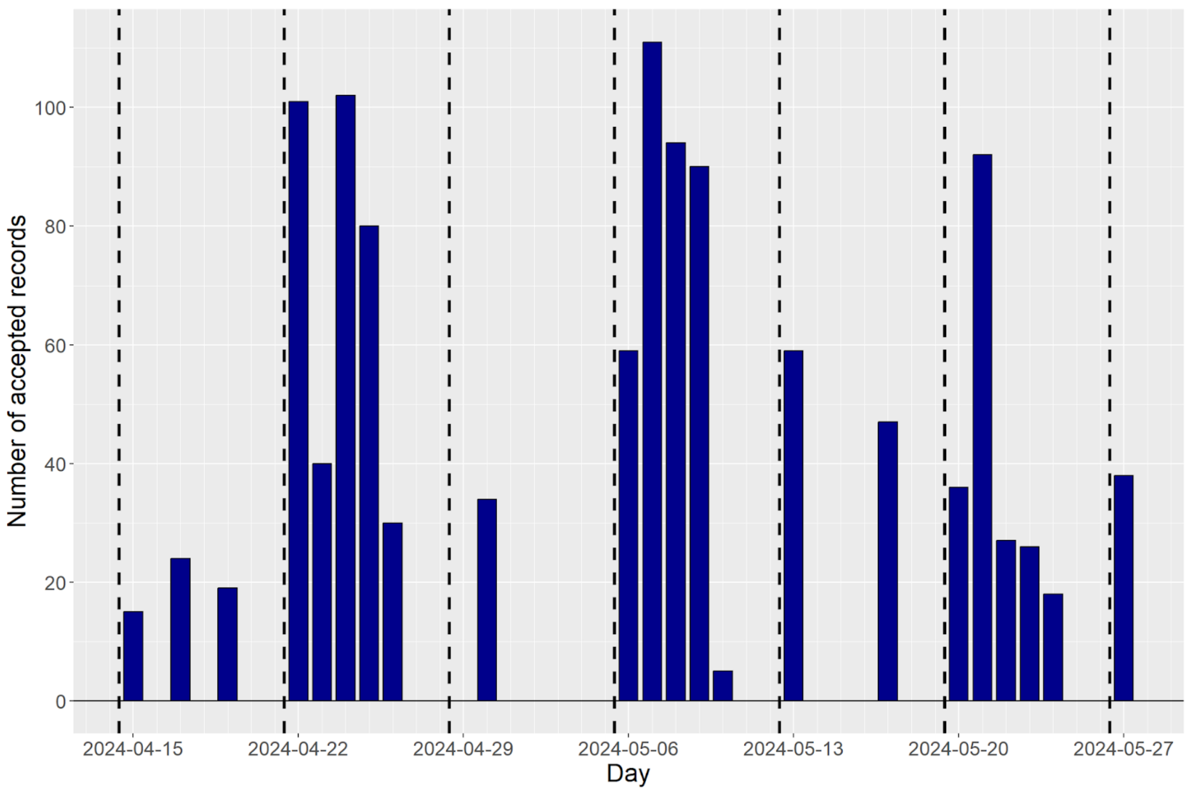


Table A3.3

Fieldwork Data Collection Flow

Date	Number of sessions	Number of students	Number of accepted records	Cumulative number of accepted records	Number of excluded records					
					Overall	NC	S	LBT	MM	T
2024-04-17	2	25	24	39	1	0	0	0	1	0
2024-04-19	2	21	19	58	2	0	0	2	0	0
2024-04-22	9	107	101	159	6	0	0	4	0	2
2024-04-23	4	48	40	199	8	0	2	4	0	2
2024-04-24	8	112	102	301	10	4	2	4	0	2
2024-04-25	8	85	80	381	5	1	3	1	1	0
2024-04-26	3	33	30	411	3	0	0	0	0	3
2024-04-30	3	38	34	445	4	0	2	0	0	2
2024-05-06	3	63	59	504	4	0	0	0	0	4
2024-05-07	6	121	111	615	10	0	2	1	1	6
2024-05-08	7	115	94	709	21	16	2	0	0	4
2024-05-09	5	93	90	799	3	1	1	0	0	2
2024-05-10	1	7	5	804	2	1	2	0	0	0
2024-05-13	5	64	59	863	5	1	2	0	0	2
2024-05-17	3	49	47	910	2	0	0	0	0	2
2024-05-20	2	40	36	946	4	3	1	0	0	0
2024-05-21	6	93	92	1038	1	1	0	0	0	0
2024-05-22	2	29	27	1065	2	1	0	0	1	0
2024-05-23	3	34	26	1091	8	0	1	0	0	7
2024-05-24	2	18	18	1109	0	0	0	0	0	0
2024-05-27	3	39	38	1147	1	1	0	0	0	0
Overall	**89**	**1250**	**1147**	**1147**	**103**	**31**	**20**	**16**	**4**	**38**

Notes:
Number of accepted records = Overall number of students – Overall number of excluded records
Reasons for excluding records:
NC – not reaching the end of the survey,
S – completing the survey in time below 10 minutes (speeding),
LBT – leaving the browser tab with the survey during the survey completion,
MM – the same login (token) was used on different computers,
T —analysis of responses and timing of responses indicates that it is highly probable that two or more students took the test together

Annex 4

Below, psychometric properties of each of the items are reported.

- GAM smooth was used to estimate nonparametric ICC, using the function `itemGAM()` from the R *mirt* package (Chalmers, 2012), with 10 plausible values drawn using the noninvariant 2PL model used as latent ability estimates.
- Reported RMSEAs are calculated using Yen's (1982) variant of a χ^2 statistic (**without** direct reference to the aforementioned GAM fit).

M032579 (item set 1)

ICCs and fit statistics

Figure A4.1.1

Item M032579 item characteristics curves (ICCs)

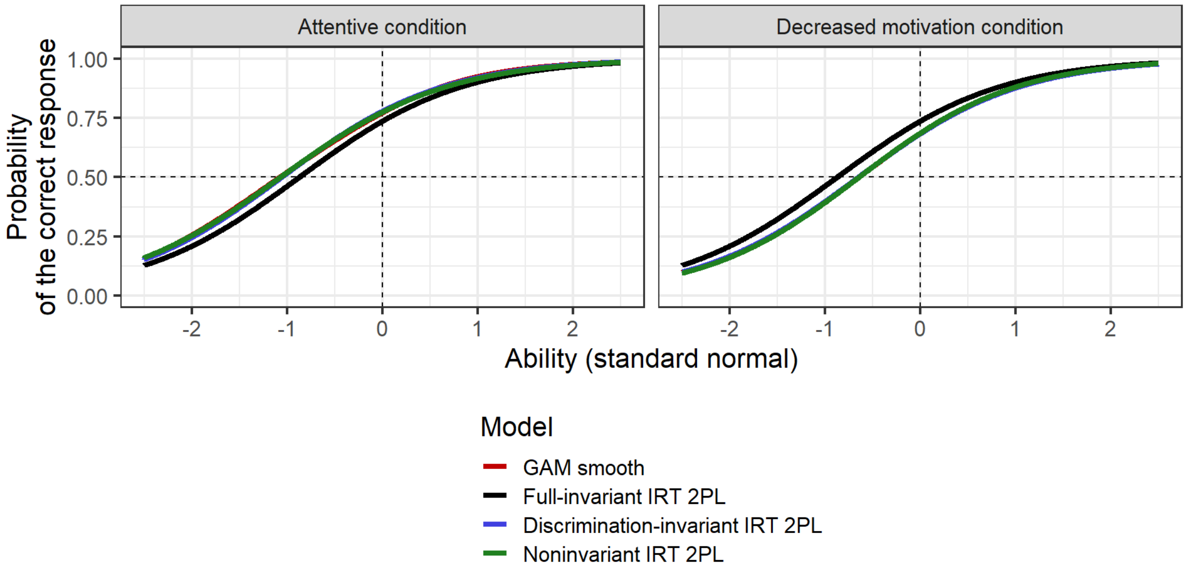


Table A4.1.1

Item M032579 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.18	−0.87		
Attentive condition			0.02	0.208
Decreased motivation condition			0.04	0.033
Discrimination-invariant 2PL model				
All groups	1.19			
Attentive condition		−1.06	0.00	0.589
Decreased motivation condition		−0.65	0.02	0.287
Noninvariant 2PL model				
Attentive condition	1.16	−1.07	0.00	0.485
Decreased motivation condition	1.22	−0.66	0.03	0.180

Frequency of answers

Figure A4.1.2

Distribution of answers to the item M032579

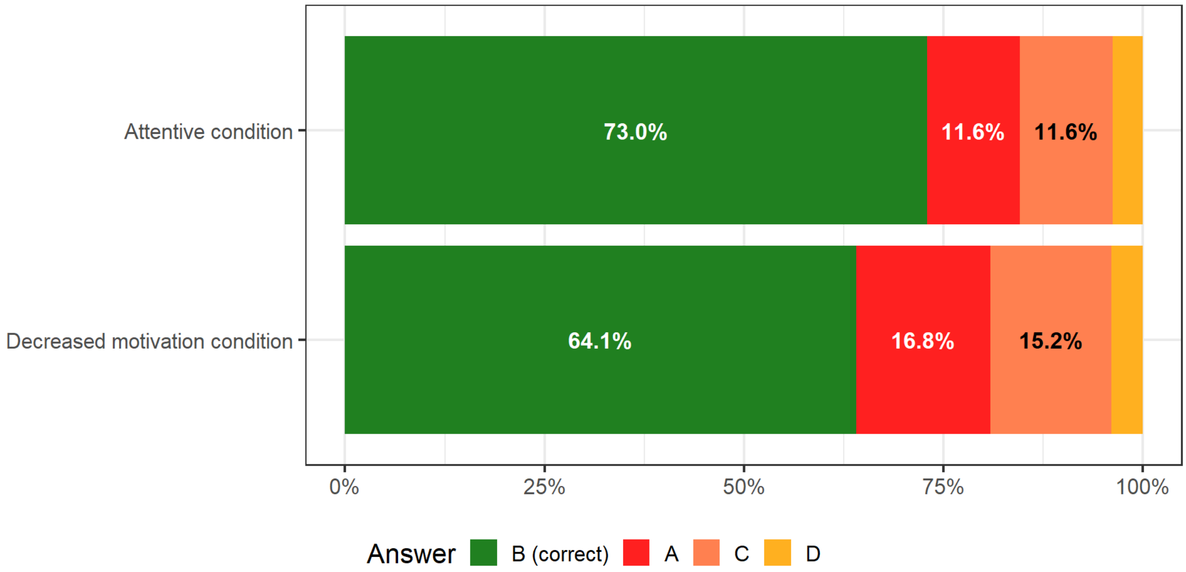


Table A4.1.2

Distribution of answers to the item M032579

Condition	**Answers**				**Overall**
	A	**B (correct)**	**C**	**D**	
Regular (attentive)	62 (11.6%)	389 (73.0%)	62 (11.6%)	20 (3.8%)	533 (100.0%)
Decreased motivation	74 (16.8%)	282 (64.1%)	67 (15.2%)	17 (3.9%)	440 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 9.76$, $p = 0.021$

Answering time

Figure A4.1.3

Distribution of response time to the item M032579

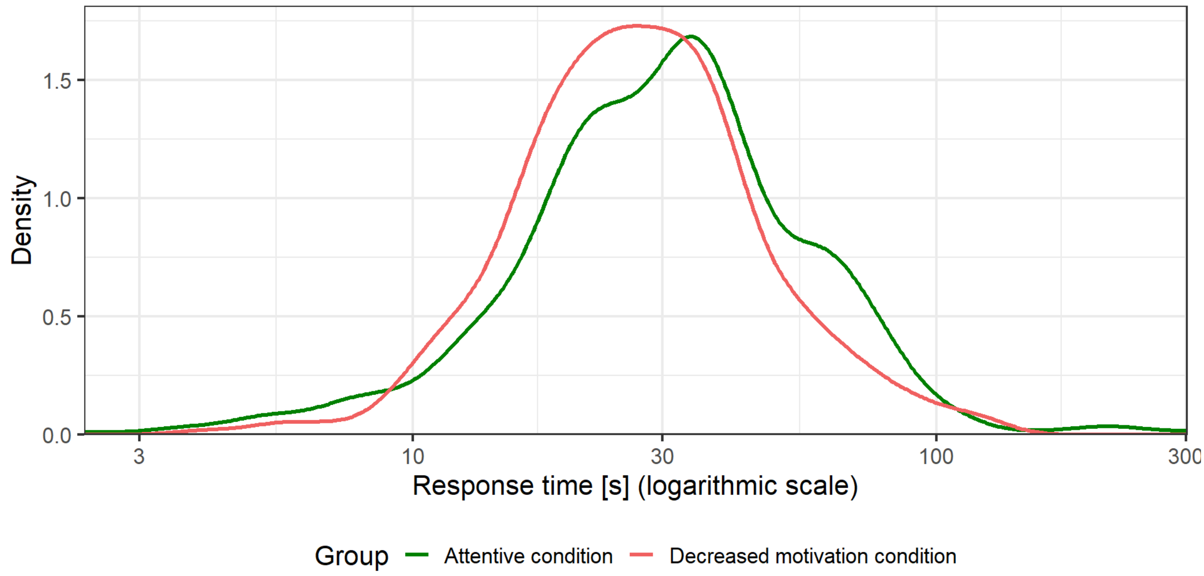


Table A4.1.3

Median and mean log response time to the item M032579

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	36.2	3.39
Decreased motivation	31.1	3.29
ANOVA for mean log(response time) differences: $F(1, 972) = 6.72$, $p = 0.010$		

M042254 (item set 1)

ICCs and fit statistics

Figure A4.2.1

Item M042254 item characteristics curves (ICCs)

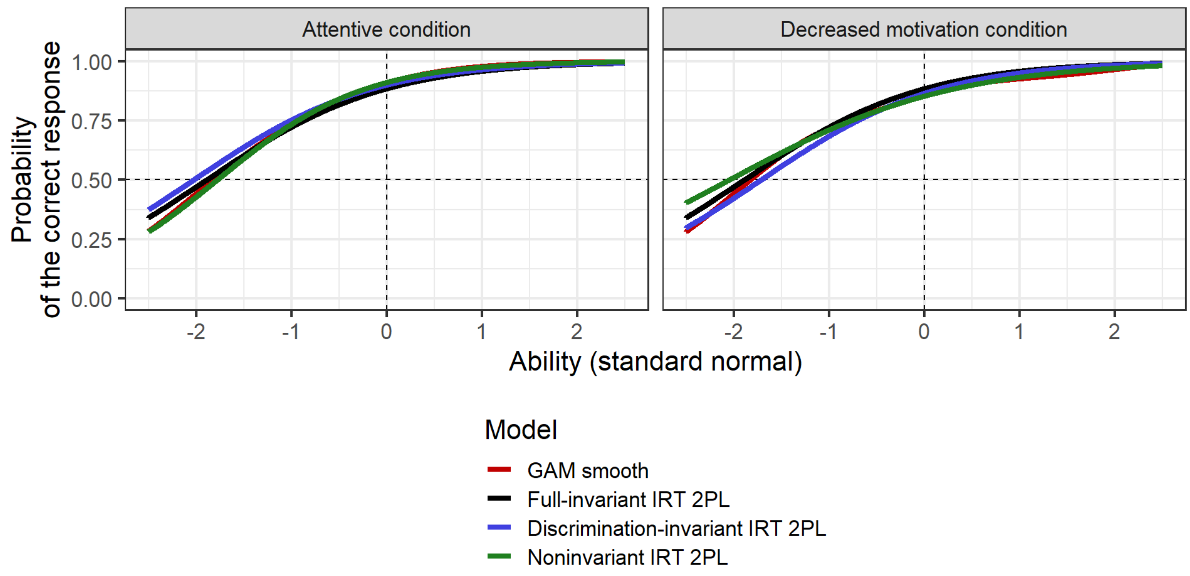


Table A4.2.1

Item M042254 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.08	−1.89		
Attentive condition			0.00	0.513
Decreased motivation condition			0.02	0.225
Discrimination-invariant 2PL model				
All groups	1.08			
Attentive condition		−2.02	0.02	0.321
Decreased motivation condition		−1.71	0.00	0.708
Noninvariant 2PL model				
Attentive condition	1.30	−1.78	0.00	0.792
Decreased motivation condition	0.86	−2.04	0.03	0.170

Frequency of answers

Figure A4.2.2

Distribution of answers to the item M042254

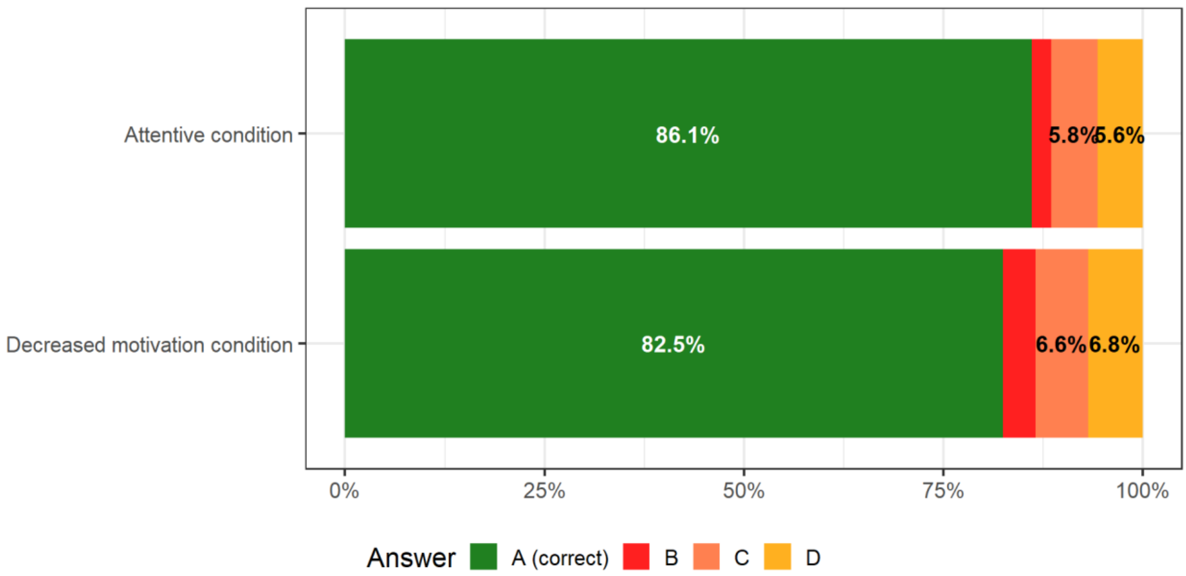


Table A4.2.2

Distribution of answers to the item M042254

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	460 (86.1%)	13 (2.4%)	31 (5.8%)	30 (5.6%)	534 (100.0%)
Decreased motivation	362 (82.5%)	18 (4.1%)	29 (6.6%)	30 (6.8%)	439 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 3.31$, $p = 0.346$

Answering time

Figure A4.2.3

Distribution of response time to the item M042254

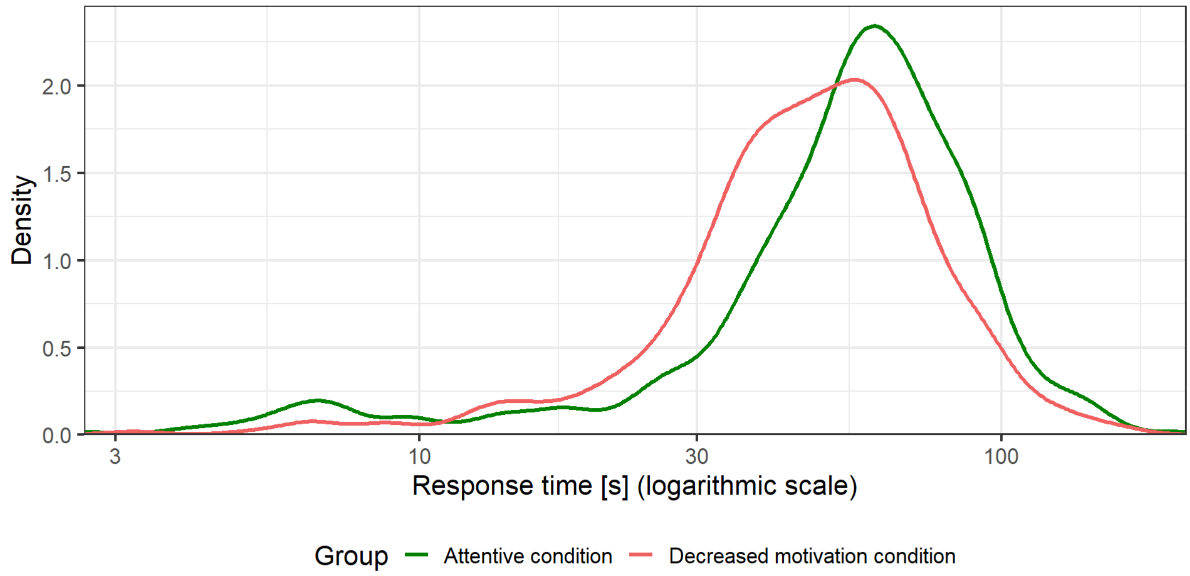


Table A4.2.3

Median and mean log response time to the item M042254

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	58.7	3.92
Decreased motivation	51.2	3.81
ANOVA for mean log(response time) differences: $F(1, 972) = 6.91$, $p = 0.009$		

M042222 (item set 1)

ICCs and fit statistics

Figure A4.3.1

Item M042222 item characteristics curves (ICCs)

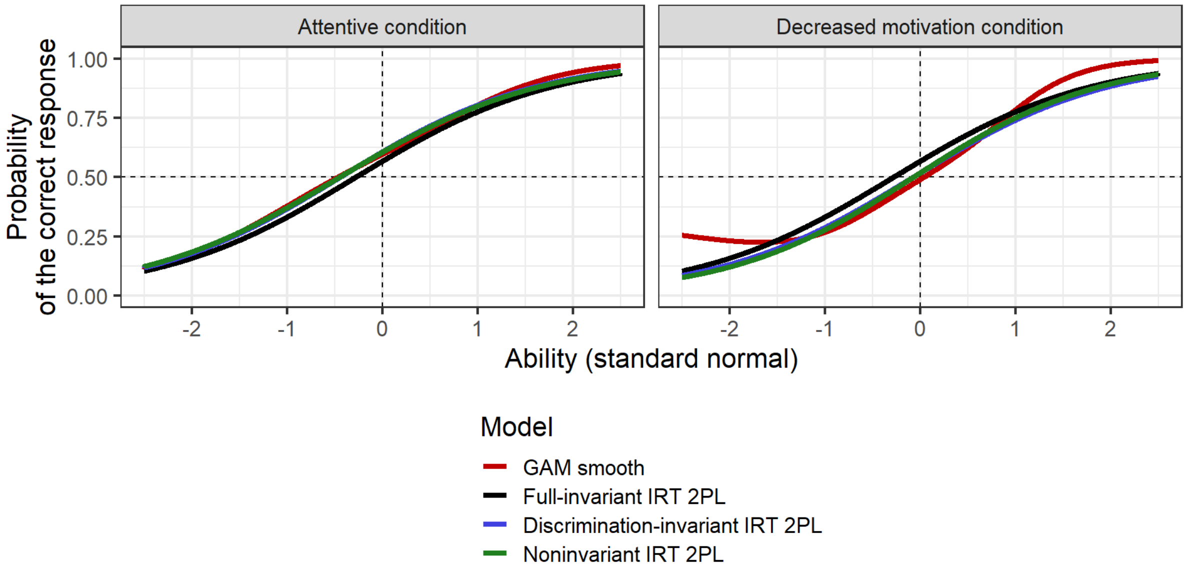


Table A4.3.1

Item M042222 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.97	−0.28		
Attentive condition			0.04	0.037
Decreased motivation condition			0.07	0.001
Discrimination-invariant 2PL model				
All groups	0.98			
Attentive condition		−0.44	0.03	0.125
Decreased motivation condition		−0.07	0.06	0.001
Noninvariant 2PL model				
Attentive condition	0.96	−0.45	0.03	0.105
Decreased motivation condition	1.03	−0.07	0.06	0.001

Frequency of answers

Figure A4.3.2

Distribution of answers to the item M042222

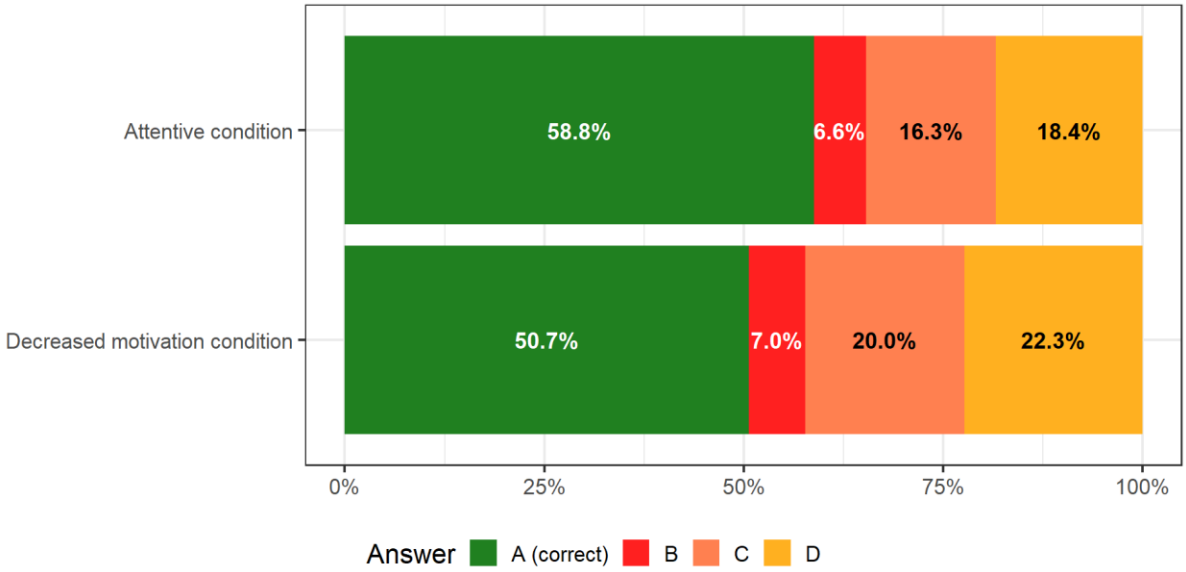


Table A4.3.2

Distribution of answers to the item M042222

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	314 (58.8%)	35 (6.6%)	87 (16.3%)	98 (18.4%)	534 (100.0%)
Decreased motivation	223 (50.7%)	31 (7.0%)	88 (20.0%)	98 (22.3%)	440 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 6.66$, $p = 0.084$

Answering time

Figure A4.3.3

Distribution of response time to the item M042222

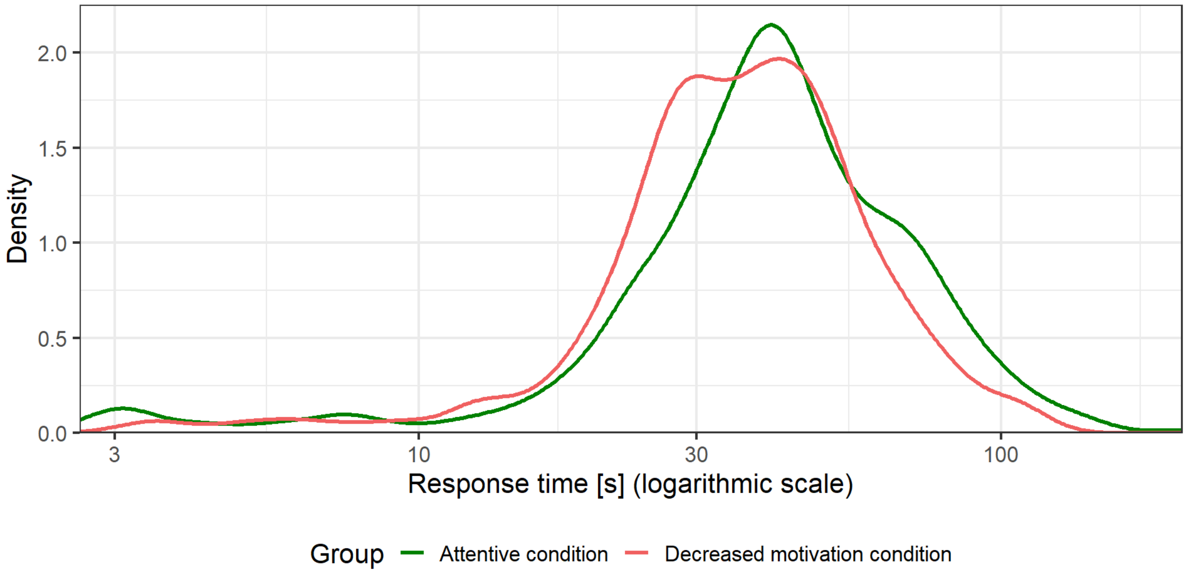


Table A4.3.3

Median and mean log response time to the item M042222

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	45.0	3.64
Decreased motivation	39.5	3.55
ANOVA for mean log(response time) differences: $F(1, 972) = 4.91$, $p = 0.027$		

gm2006z08 (item set 1)

ICCs and fit statistics

Figure A4.4.1

Item gm2006z08 item characteristics curves (ICCs)

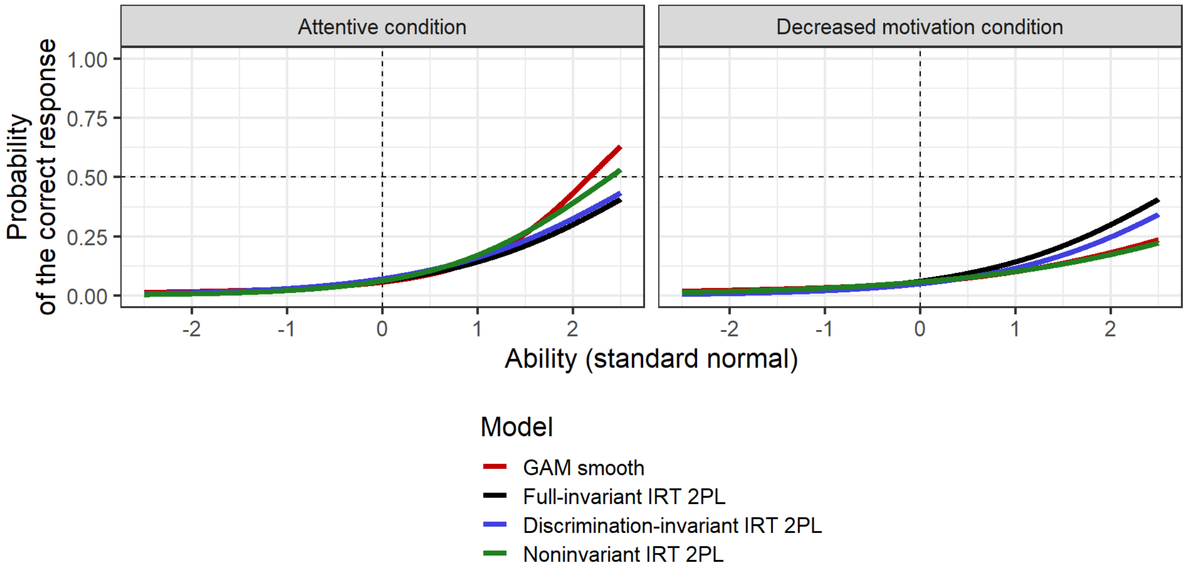


Table A4.4.1

Item gm2006z08 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.95	2.90−		
Attentive condition			0.06	0.001
Decreased motivation condition			0.04	0.087
Discrimination-invariant 2PL model				
All groups	0.92			
Attentive condition		2.79	0.06	0.001
Decreased motivation condition		3.20	0.04	0.063
Noninvariant 2PL model				
Attentive condition	1.14	2.39	0.06	0.002
Decreased motivation condition	0.62	4.53	0.02	0.237

Frequency of answers

Figure A4.4.2

Distribution of answers to the item gm2006z08

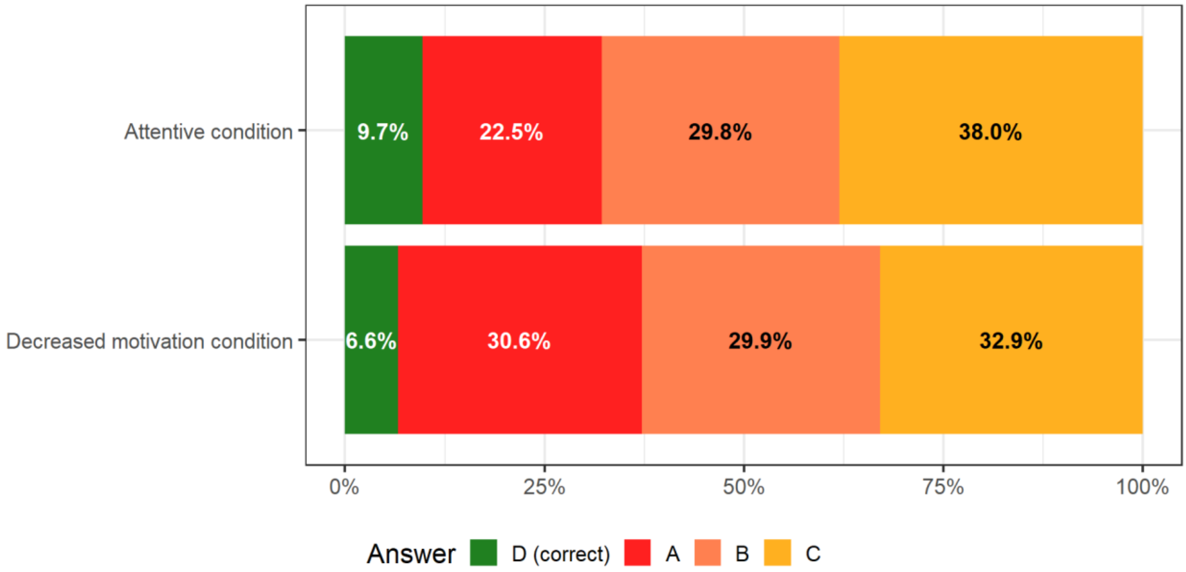


Table A4.4.2

Distribution of answers to the item gm2006z08

Condition	Answers				Overall
	A	B	C	D (correct)	
Regular (attentive)	120 (22.5%)	159 (29.8%)	203 (38.0%)	52 (9.7%)	534 (100.0%)
Decreased motivation	134 (30.6%)	131 (29.9%)	144 (32.9%)	29 (6.6%)	438 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 10.66$, $p = 0.014$

Answering time

Figure A4.4.3

Distribution of response time to the item gm2006z08

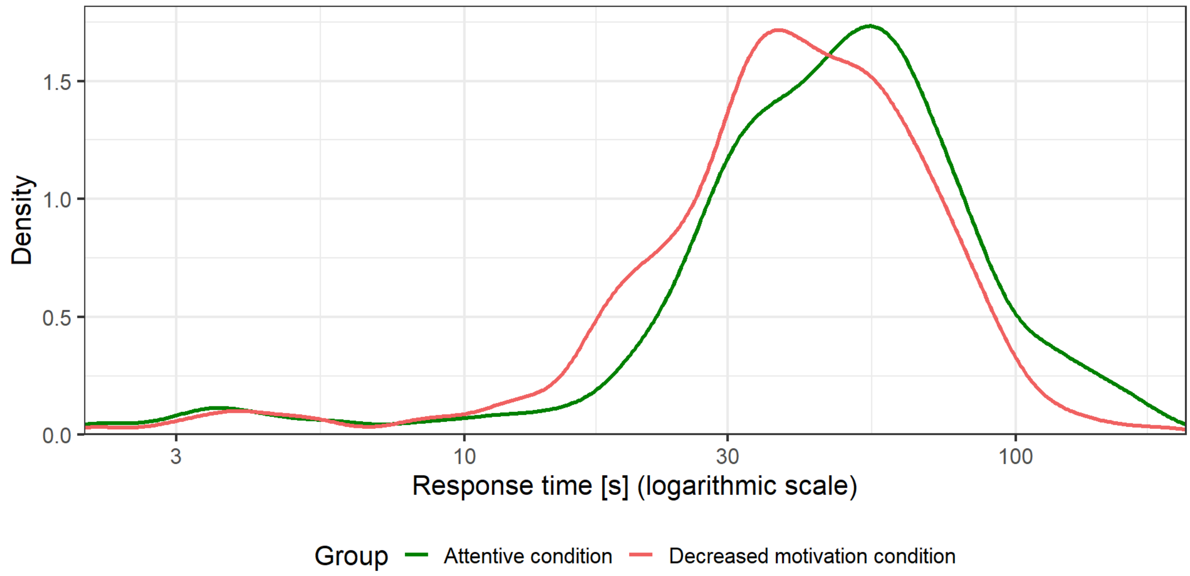


Table A4.4.3

Median and mean log response time to the item gm2006z08

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	52.6	3.76
Decreased motivation	44.2	3.61
ANOVA for mean log(response time) differences: $F(1, 972) = 10.44$, $p = 0.001$		

M042260 (item set 1)

ICCs and fit statistics

Figure A4.5.1

Item M042260 item characteristics curves (ICCs)

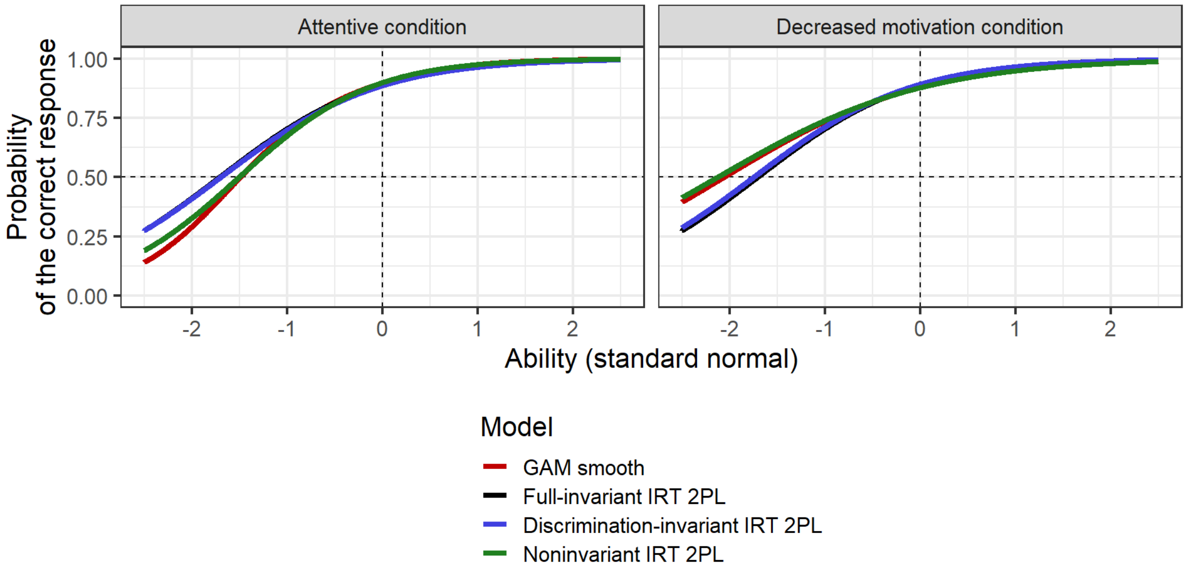


Table A4.5.1

Item M042260 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.22	−1.71		
Attentive condition			0.03	0.186
Decreased motivation condition			0.00	0.852
Discrimination-invariant 2PL model				
All groups	1.21			
Attentive condition		−1.69	0.01	0.398
Decreased motivation condition		−1.74	0.00	0.958
Noninvariant 2PL model				
Attentive condition	1.45	−1.50	0.00	0.798
Decreased motivation condition	0.93	−2.12	0.00	0.615

Frequency of answers

Figure A4.5.2

Distribution of answers to the item M042260

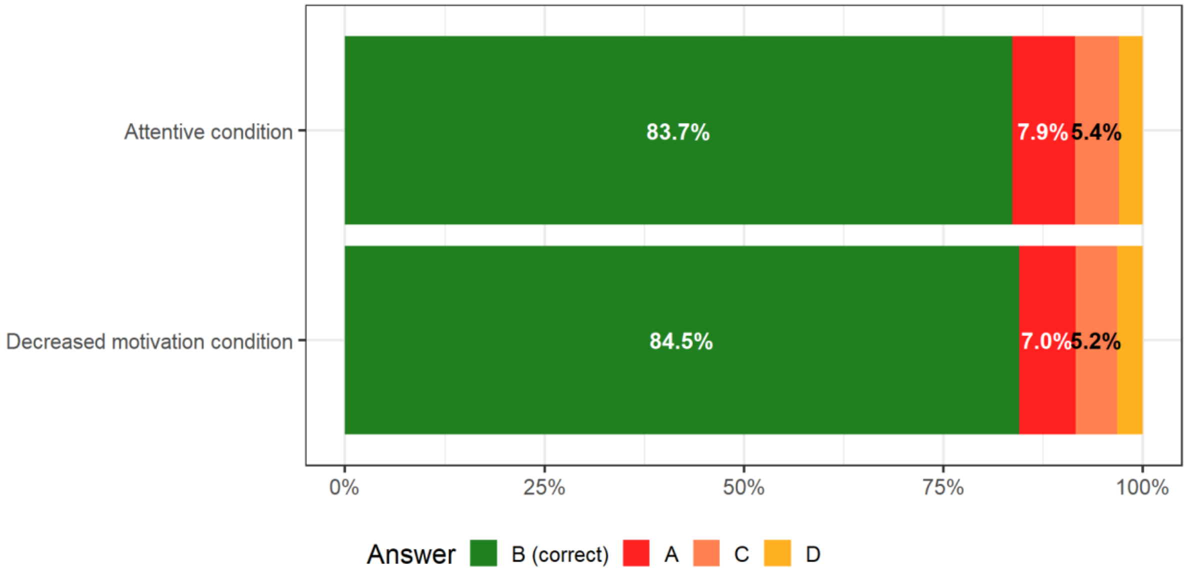


Table A4.5.2

Distribution of answers to the item M042260

Condition	**Answers**				**Overall**
	A	**B (correct)**	**C**	**D**	
Regular (attentive)	42 (7.9%)	447 (83.7%)	29 (5.4%)	16 (3.0%)	534 (100.0%)
Decreased motivation	31 (7.0%)	372 (84.5%)	23 (5.2%)	14 (3.2%)	440 (100.0%)
Pearson Chi-squared tests of independence: $\chi^2(3) = 0.28$, $p = 0.963$					

Answering time

Figure A4.5.3

Distribution of response time to the item M042260

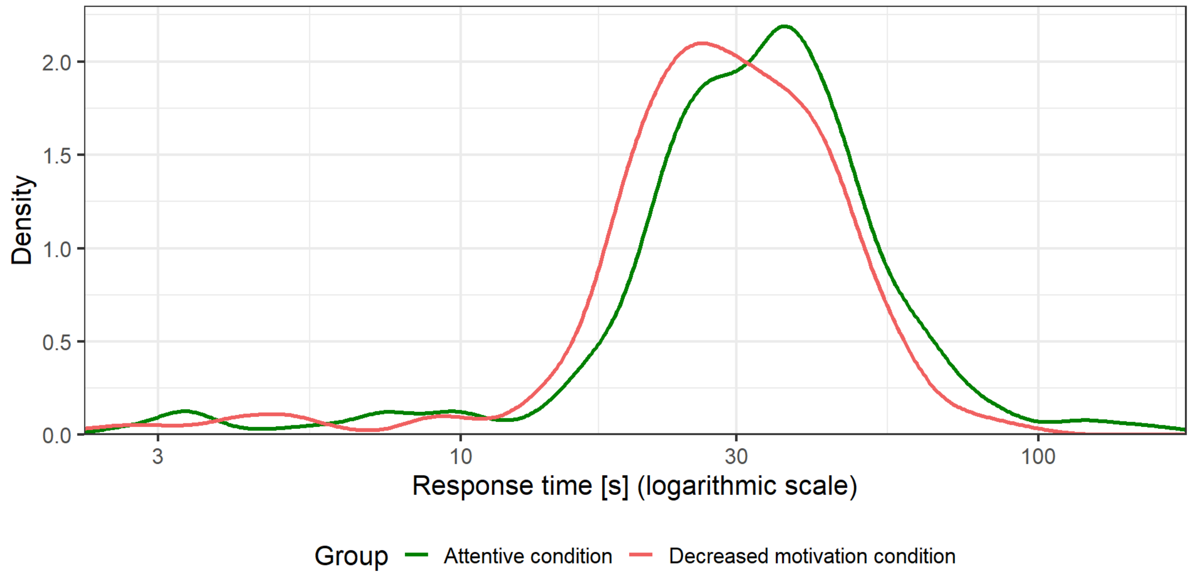


Table A4.5.3

Median and mean log response time to the item M042260

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	35.8	3.43
Decreased motivation	30.9	3.32

ANOVA for mean log(response time) differences: $F(1, 972) = 10.59$, $p = 0.001$

gmm2015z07 (item set 1)

ICCs and fit statistics

Figure A4.6.1

Item gmm2015z07 item characteristics curves (ICCs)

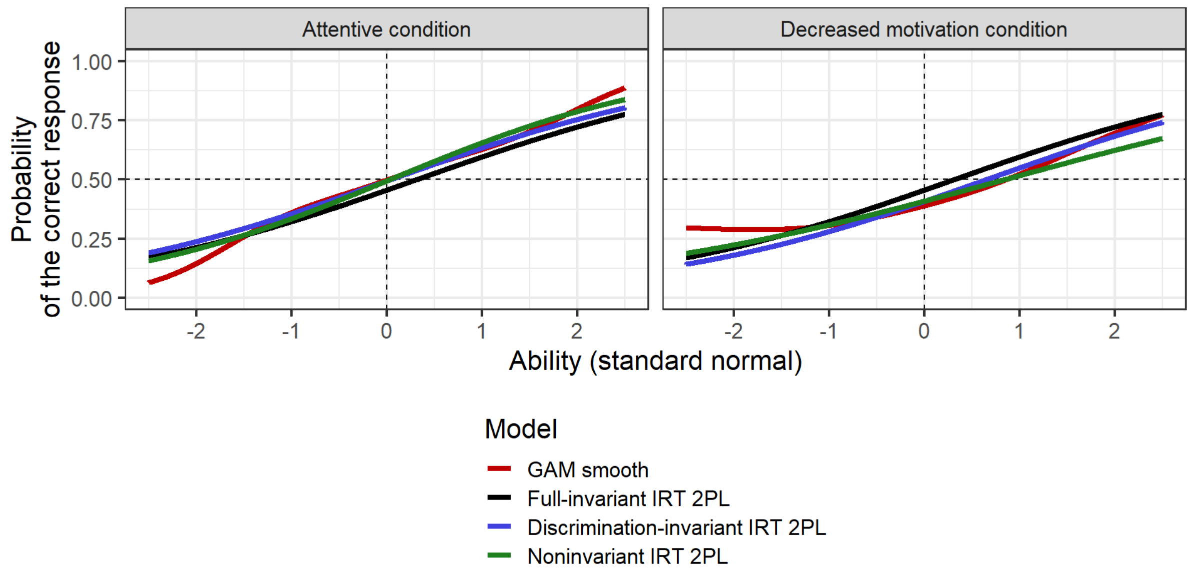


Table A4.6.1

Item gmm2015z07 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.57	0.32		
Attentive condition			0.05	0.018
Decreased motivation condition			0.04	0.049
Discrimination-invariant 2PL model				
All groups	0.57			
Attentive condition		0.04	0.04	0.044
Decreased motivation condition		0.66	0.03	0.125
Noninvariant 2PL model				
Attentive condition	0.66	0.04	0.02	0.299
Decreased motivation condition	0.44	0.85	0.03	0.122

Frequency of answers

Figure A4.6.2

Distribution of answers to the item gmm2015z07

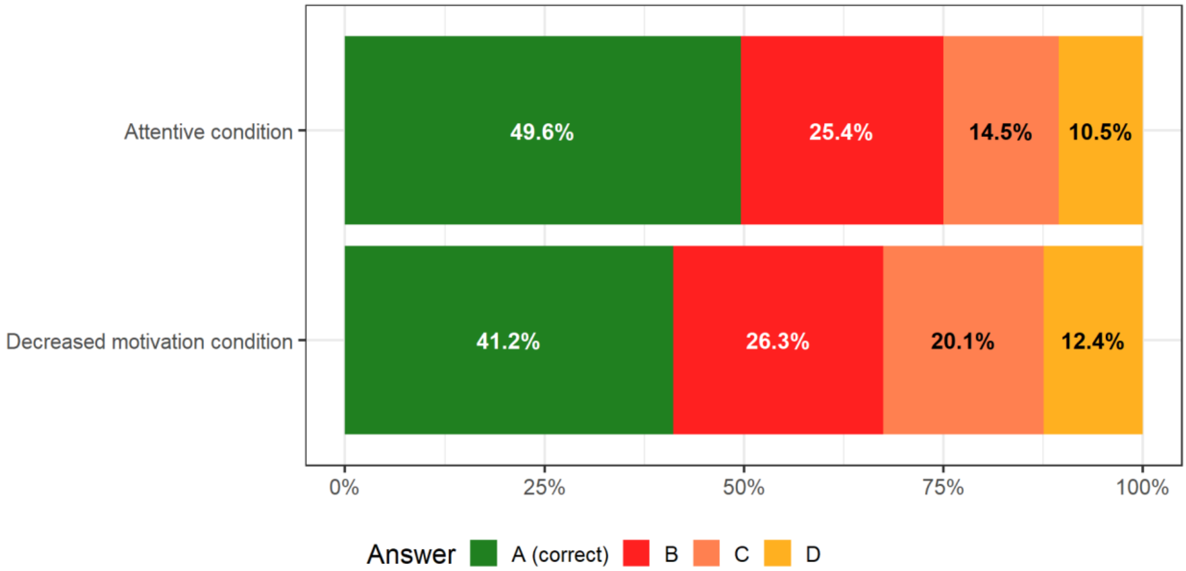


Table A4.6.2

Distribution of answers to the item gmm2015z07

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	264 (49.6%)	135 (25.4%)	77 (14.5%)	56 (10.5%)	532 (100.0%)
Decreased motivation	180 (41.2%)	115 (26.3%)	88 (20.1%)	54 (12.4%)	437 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 9.03$, $p = 0.029$

Answering time

Figure A4.6.3

Distribution of response time to the item gmm2015z07

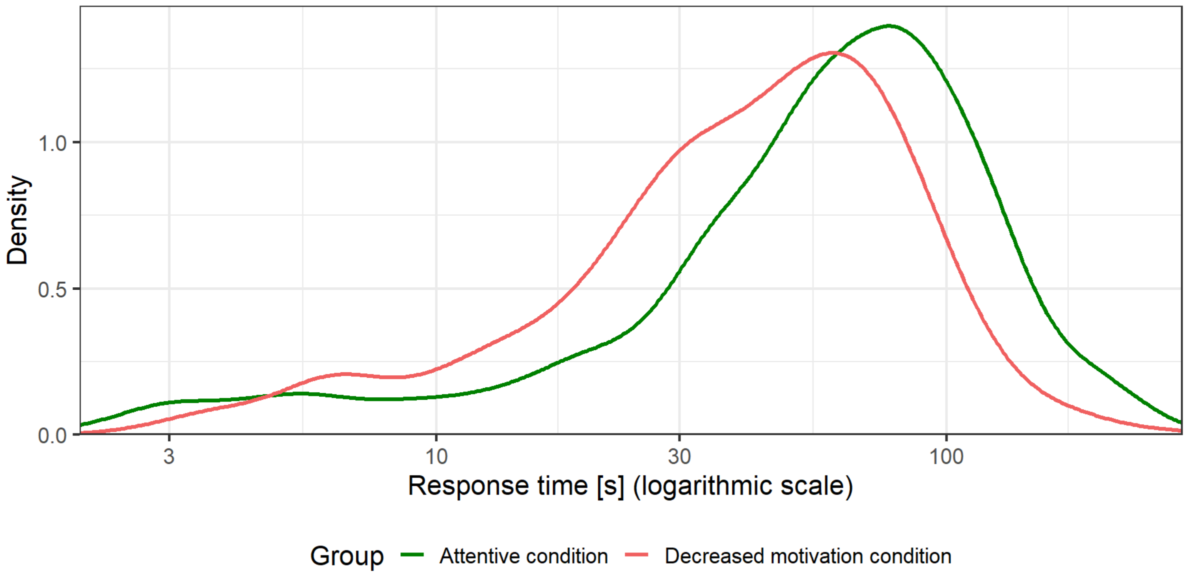


Table A4.6.3

Median and mean log response time to the item gmm2015z07

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	68.2	3.92
Decreased motivation	49.5	3.63

ANOVA for mean log(response time) differences: $F(1, 972) = 25.35$, $p = 0.000$

M032295 (item set 1)

ICCs and fit statistics

Figure A4.7.1

Item M032295 item characteristics curves (ICCs)

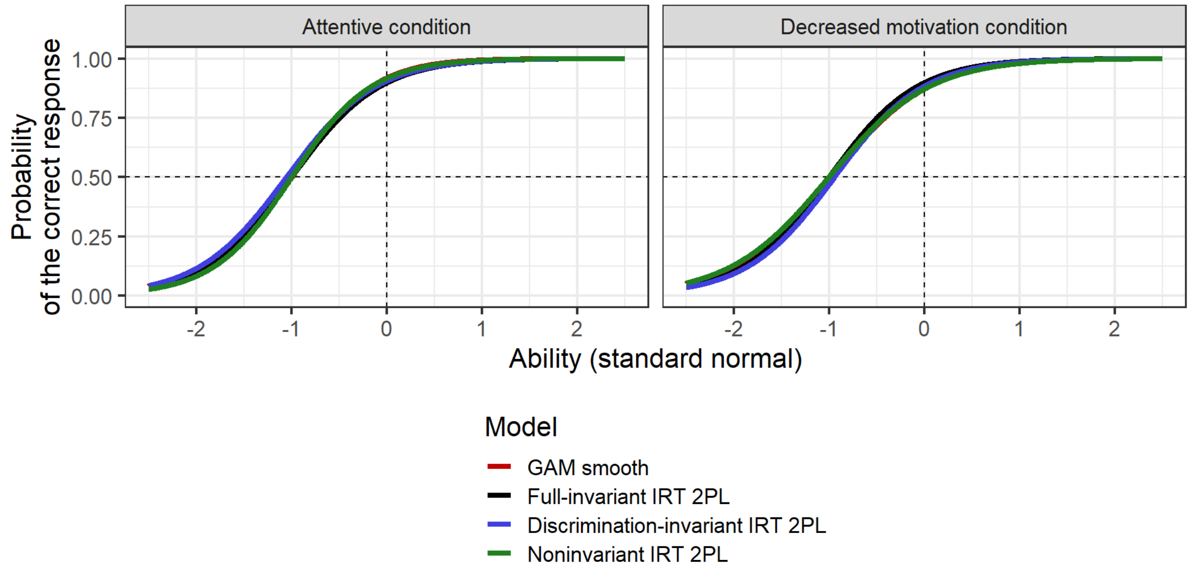


Table A4.7.1

Item M032295 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	2.18	−1.00		
Attentive condition			0.04	0.039
Decreased motivation condition			0.02	0.325
Discrimination-invariant 2PL model				
All groups	2.16			
Attentive condition		−1.05	0.05	0.010
Decreased motivation condition		−0.94	0.01	0.355
Noninvariant 2PL model				
Attentive condition	2.41	−1.00	0.04	0.036
Decreased motivation condition	1.92	−1.00	0.02	0.309

Frequency of answers

Figure A4.7.2

Distribution of answers to the item M032295

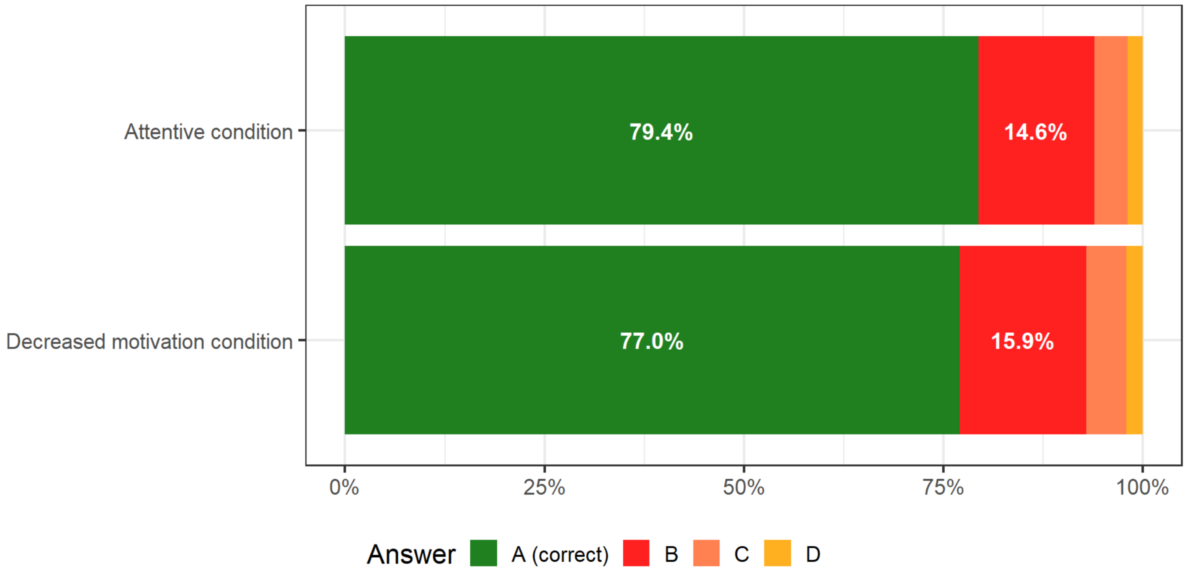


Table A4.7.2

Distribution of answers to the item M032295

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	424 (79.4%)	78 (14.6%)	22 (4.1%)	10 (1.9%)	534 (100.0%)
Decreased motivation	339 (77.0%)	70 (15.9%)	22 (5.0%)	9 (2.0%)	440 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 0.89$, $p = 0.828$

Answering time

Figure A4.7.3

Distribution of response time to the item M032295

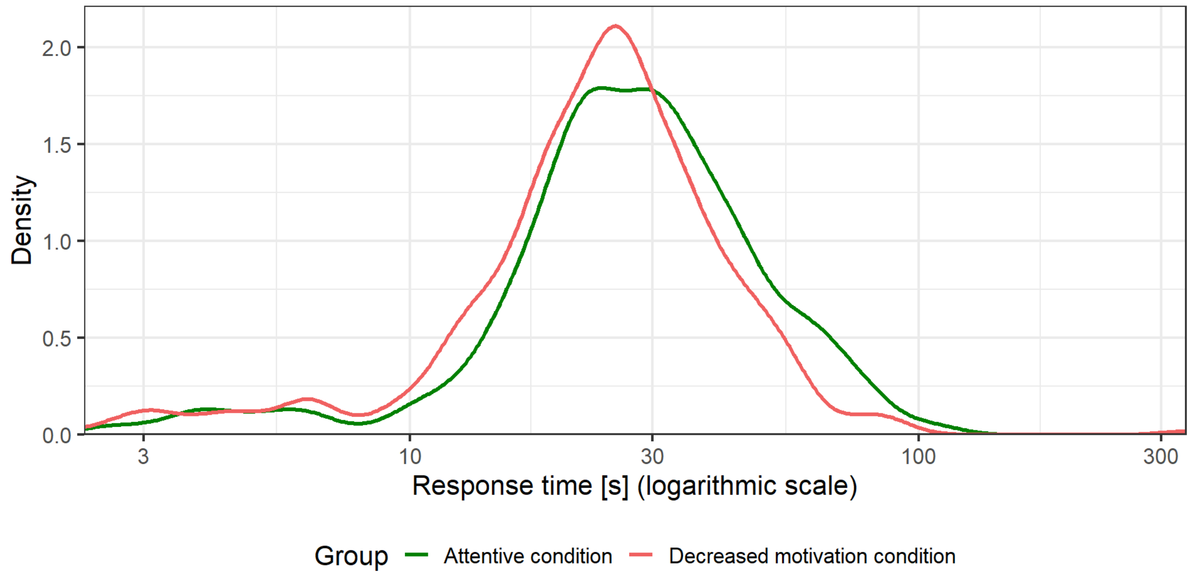


Table A4.7.3

Median and mean log response time to the item M032295

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	31.2	3.28
Decreased motivation	27.2	3.13
ANOVA for mean log(response time) differences: $F(1, 972) = 13.72$, $p = 0.000$		

M042031 (item set 1)

ICCs and fit statistics

Figure A4.8.1

Item M042031 item characteristics curves (ICCs)

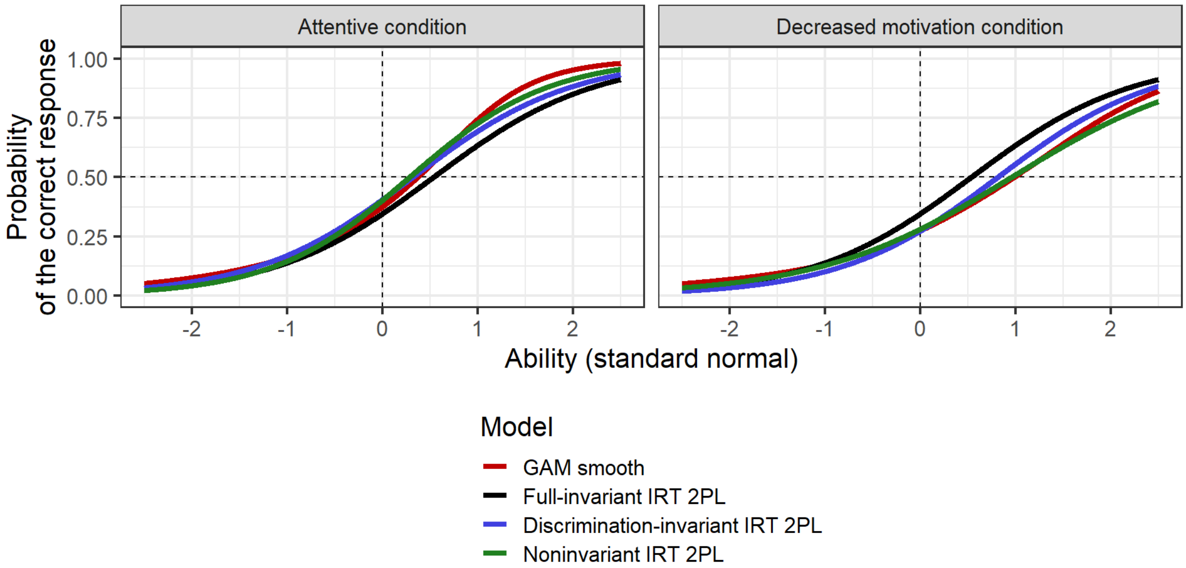


Table A4.8.1

Item M042031 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.19	0.54		
Attentive condition			0.07	0.000
Decreased motivation condition			0.04	0.056
Discrimination-invariant 2PL model				
All groups	1.20			
Attentive condition		0.32	0.06	0.002
Decreased motivation condition		0.82	0.02	0.326
Noninvariant 2PL model				
Attentive condition	1.38	0.29	0.05	0.010
Decreased motivation condition	0.98	0.96	0.03	0.165

Frequency of answers

Figure A4.8.2

Distribution of answers to the item M042031

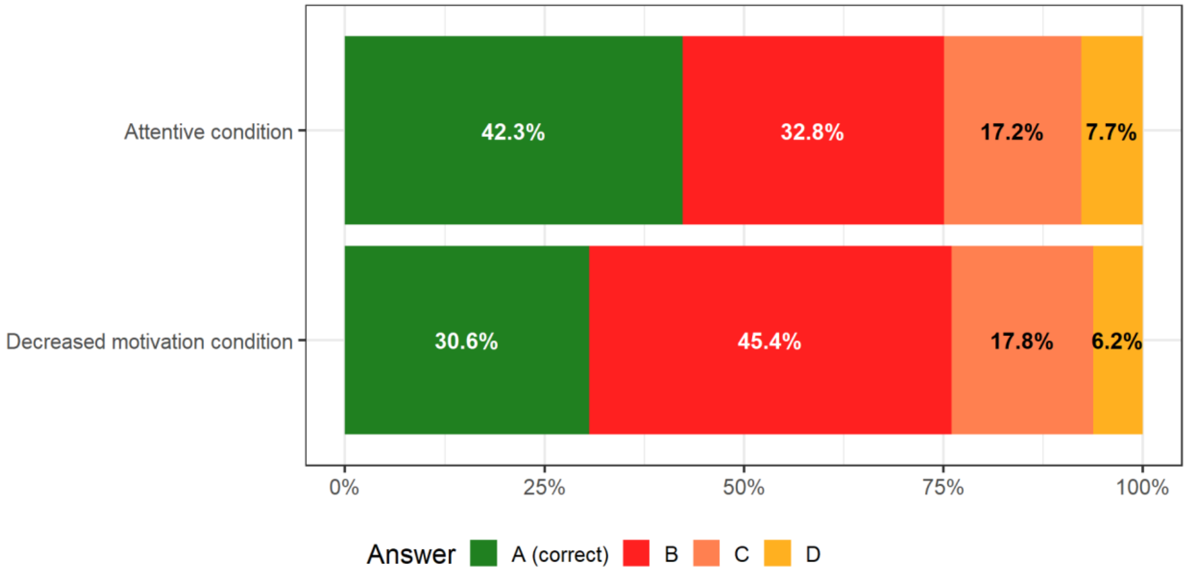


Table A4.8.2

Distribution of answers to the item M042031

Condition	**Answers**				**Overall**
	A (correct)	**B**	**C**	**D**	
Regular (attentive)	226 (42.3%)	175 (32.8%)	92 (17.2%)	41 (7.7%)	534 (100.0%)
Decreased motivation	134 (30.6%)	199 (45.4%)	78 (17.8%)	27 (6.2%)	438 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 19.80$, $p = 0.000$

Answering time

Figure A4.8.3

Distribution of response time to the item M042031

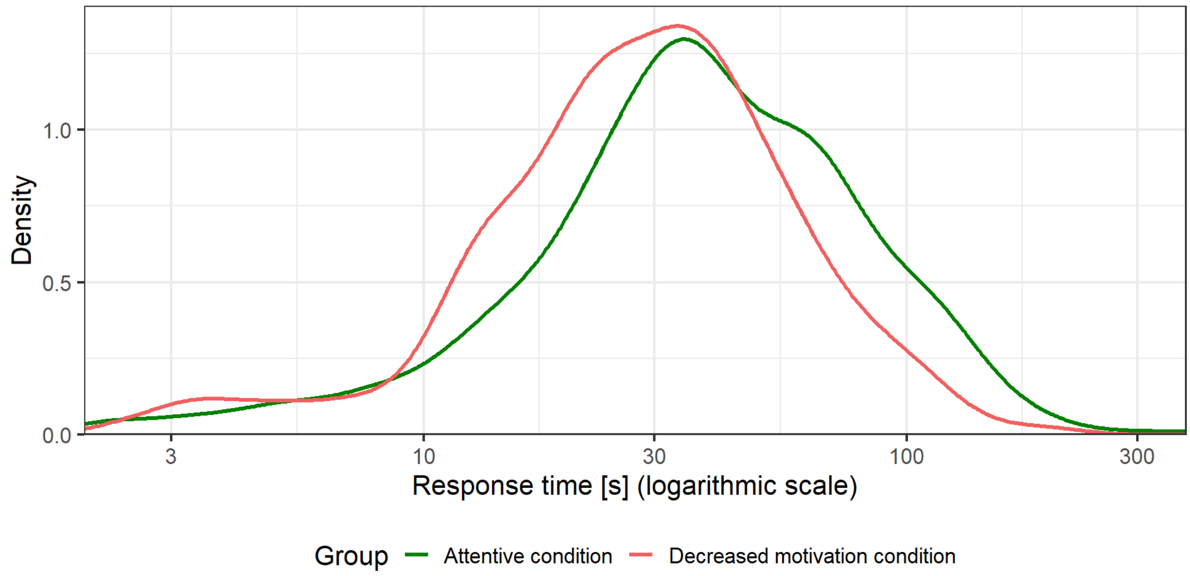


Table A4.8.3

Median and mean log response time to the item M042031

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	47.3	3.56
Decreased motivation	35.3	3.32
ANOVA for mean log(response time) differences: *F*(1, 972) = 23.70, *p* = 0.000		

gm2004z04 (item set 1)

ICCs and fit statistics

Figure A4.9.1

Item gm2004z04 item characteristics curves (ICCs)

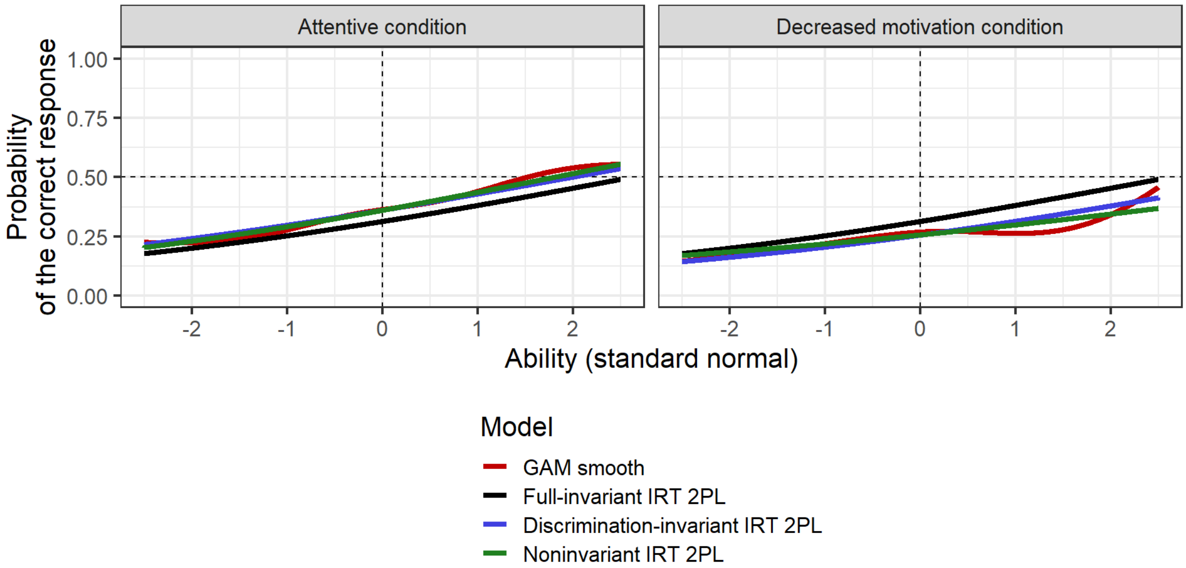


Table A4.9.1

Item gm2004z04 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.30	2.63		
Attentive condition			0.03	0.092
Decreased motivation condition			0.03	0.173
Discrimination-invariant 2PL model				
All groups	0.29			
Attentive condition		2.00	0.04	0.033
Decreased motivation condition		3.73	0.00	0.691
Noninvariant 2PL model				
Attentive condition	0.32	1.81	0.04	0.052
Decreased motivation condition	0.21	5.06	0.00	0.700

Frequency of answers

Figure A4.9.2

Distribution of answers to the item gm2004z04

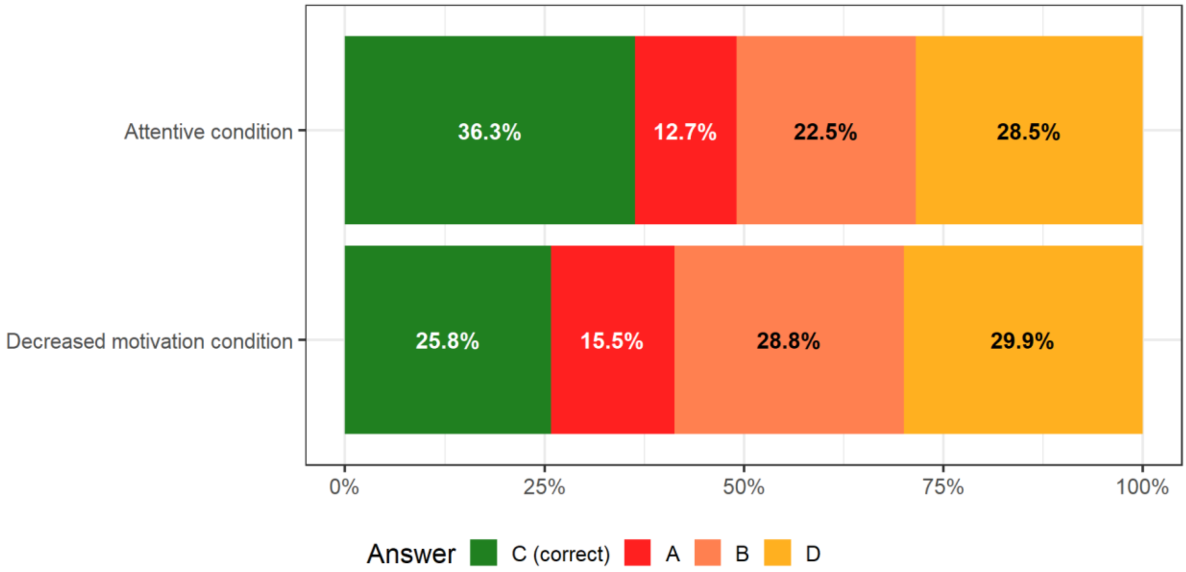


Table A4.9.2

Distribution of answers to the item gm2004z04

Condition	**Answers**				**Overall**
	A	**B**	**C (correct)**	**D**	
Regular (attentive)	68 (12.7%)	120 (22.5%)	194 (36.3%)	152 (28.5%)	534 (100.0%)
Decreased motivation	68 (15.5%)	126 (28.8%)	113 (25.8%)	131 (29.9%)	438 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 13.73$, $p = 0.003$

Answering time

Figure A4.9.3

Distribution of response time to the item gm2004z04

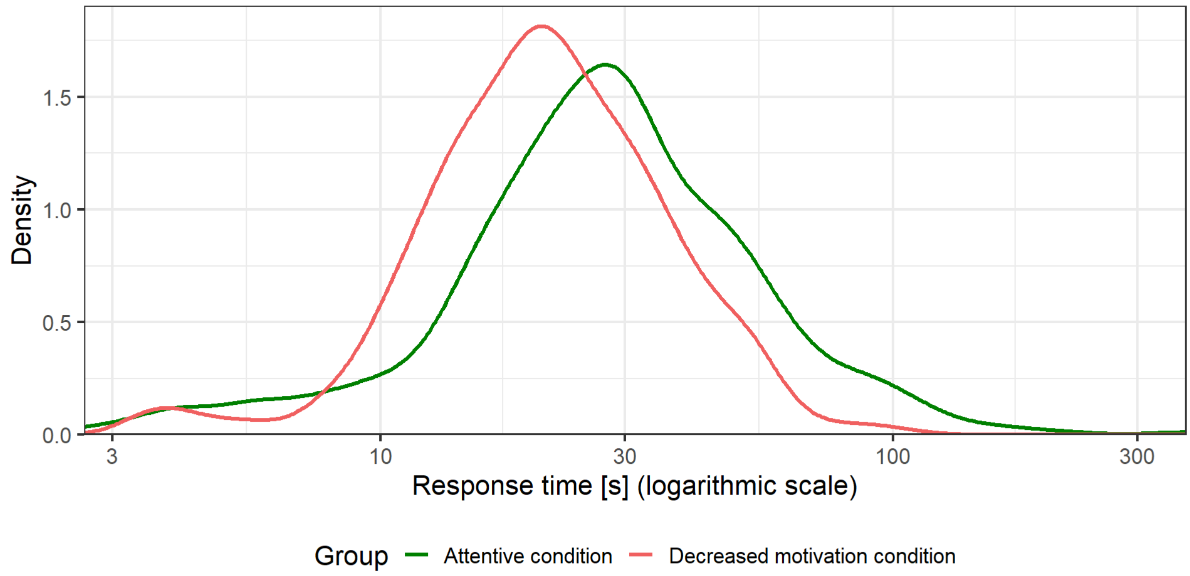


Table A4.9.3

Median and mean log response time to the item gm2004z04

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	33.8	3.28
Decreased motivation	24.0	3.04
ANOVA for mean log(response time) differences: $F(1, 972) = 35.99$, $p = 0.000$		

M032679 (item set 1)

ICCs and fit statistics

Figure A4.10.1

Item M032679 item characteristics curves (ICCs)

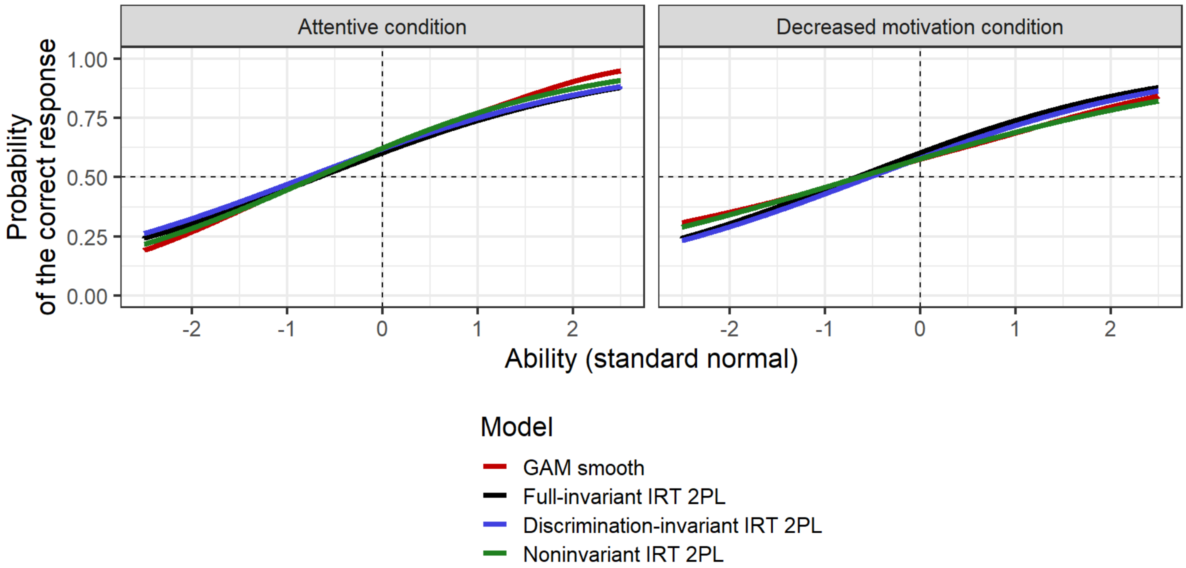


Table A4.10.1

Item M032679 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.62	−0.67		
Attentive condition			0.04	0.058
Decreased motivation condition			0.01	0.413
Discrimination-invariant 2PL model				
All groups	0.61			
Attentive condition		−0.80	0.02	0.226
Decreased motivation condition		−0.53	0.00	0.613
Noninvariant 2PL model				
Attentive condition	0.71	−0.70	0.02	0.225
Decreased motivation condition	0.48	−0.64	0.00	0.725

Frequency of answers

Figure A4.10.2

Distribution of answers to the item M032679

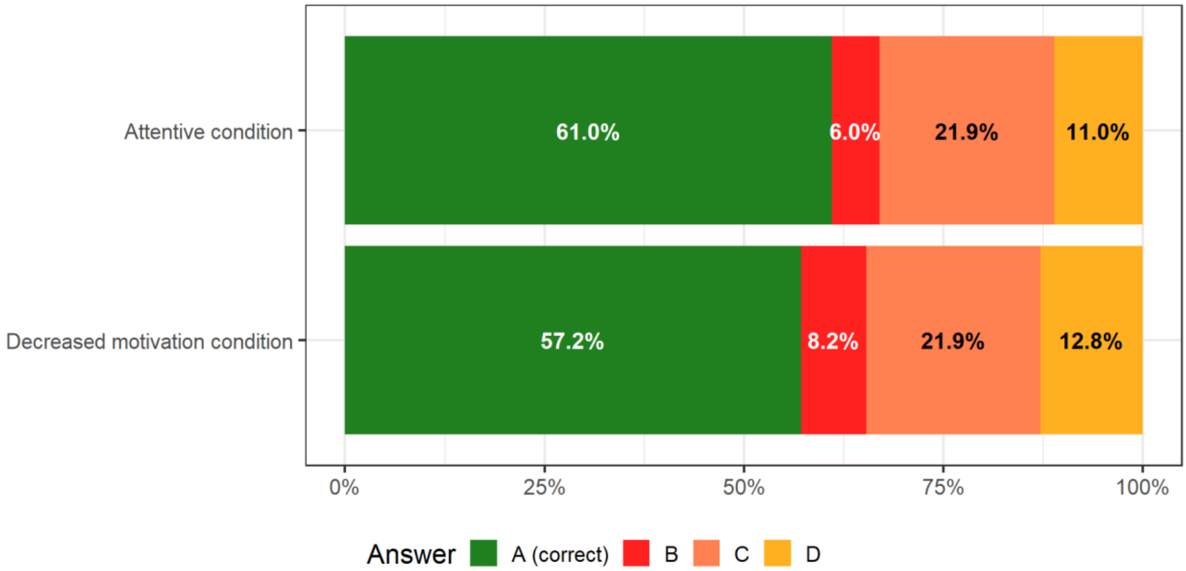


Table A4.10.2

Distribution of answers to the item M032679

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	326 (61.0%)	32 (6.0%)	117 (21.9%)	59 (11.0%)	534 (100.0%)
Decreased motivation	251 (57.2%)	36 (8.2%)	96 (21.9%)	56 (12.8%)	439 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 2.88$, $p = 0.410$

Answering time

Figure A4.10.3

Distribution of response time to the item M032679

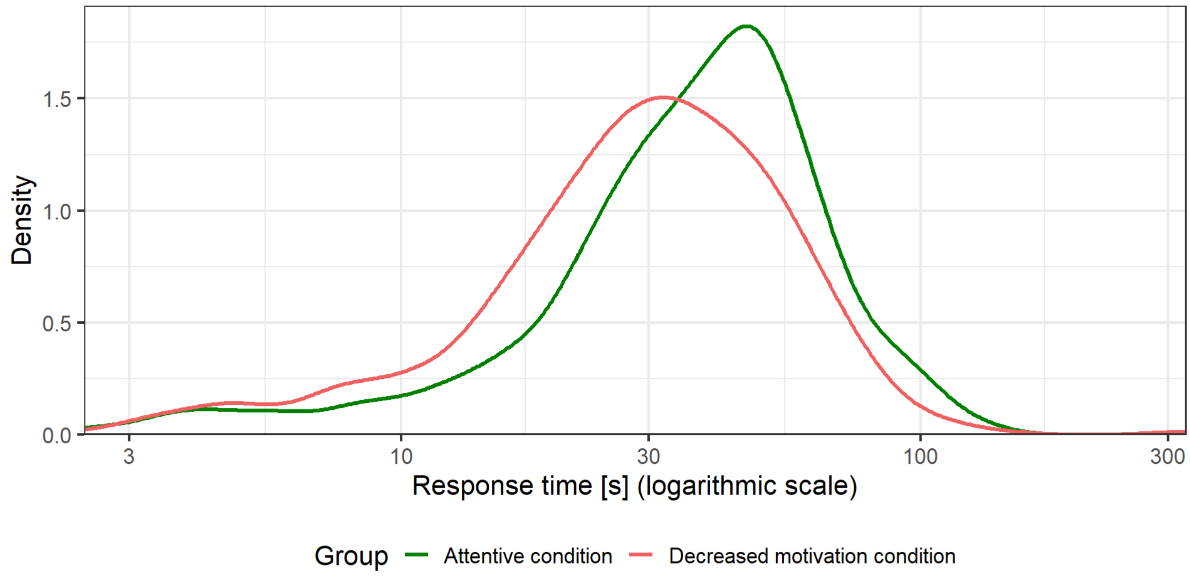


Table A4.10.3

Median and mean log response time to the item M032679

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	40.7	3.52
Decreased motivation	34.3	3.32
ANOVA for mean log(response time) differences: *F*(1, 972) = 19.87, *p* = 0.000		

M032198 (item set 1)

ICCs and fit statistics

Figure A4.11.1

Item M032198 item characteristics curves (ICCs)

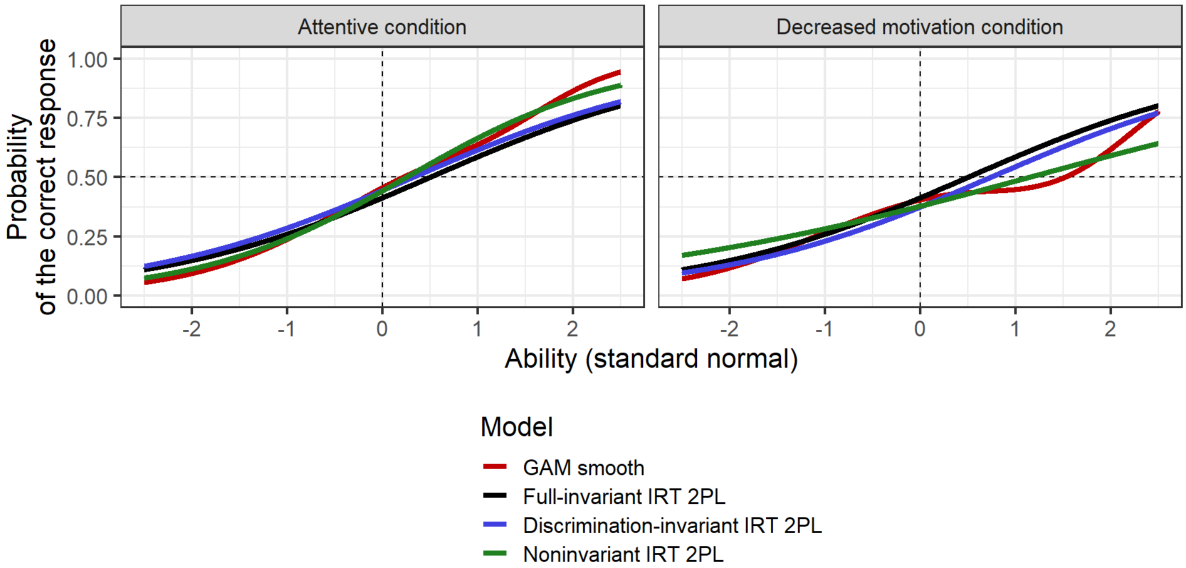


Table A4.11.1

Item M032198 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.70	0.50		
Attentive condition			0.07	0.000
Decreased motivation condition			0.02	0.262
Discrimination-invariant 2PL model				
All groups	0.69			
Attentive condition		0.32	0.05	0.015
Decreased motivation condition		0.74	0.02	0.227
Noninvariant 2PL model				
Attentive condition	0.92	0.25	0.04	0.021
Decreased motivation condition	0.43	1.15	0.03	0.093

Frequency of answers

Figure A4.11.2

Distribution of answers to the item M032198

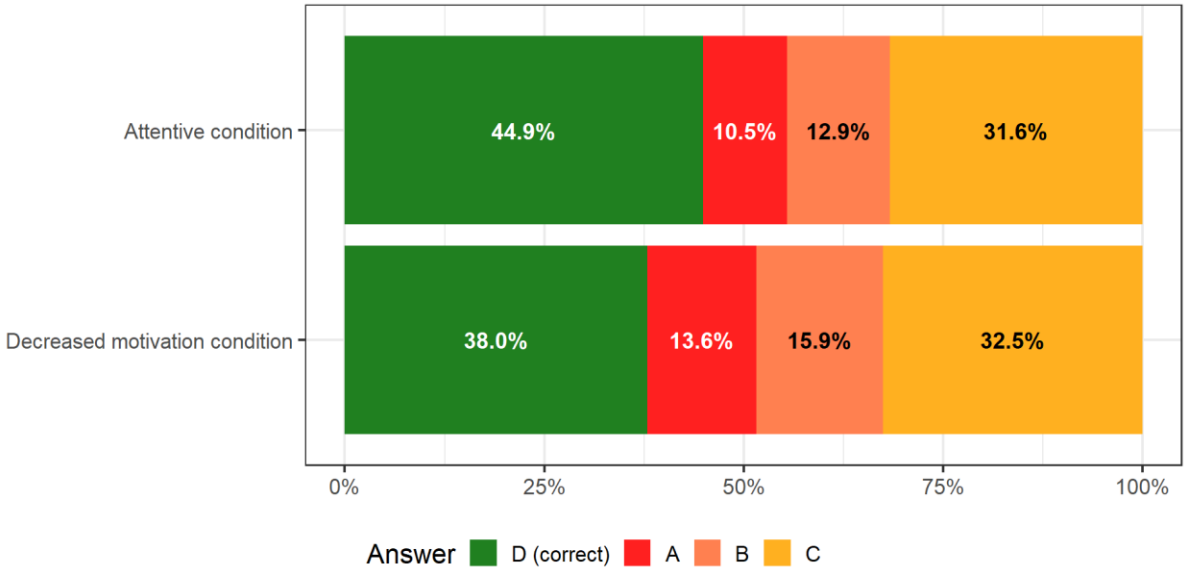


Table A4.11.2

Distribution of answers to the item M032198

Condition	Answers				Overall
	A	B	C	D (correct)	
Regular (attentive)	56 (10.5%)	69 (12.9%)	169 (31.6%)	240 (44.9%)	534 (100.0%)
Decreased motivation	60 (13.6%)	70 (15.9%)	143 (32.5%)	167 (38.0%)	440 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 6.39$, $p = 0.094$

Answering time

Figure A4.11.3

Distribution of response time to the item M032198

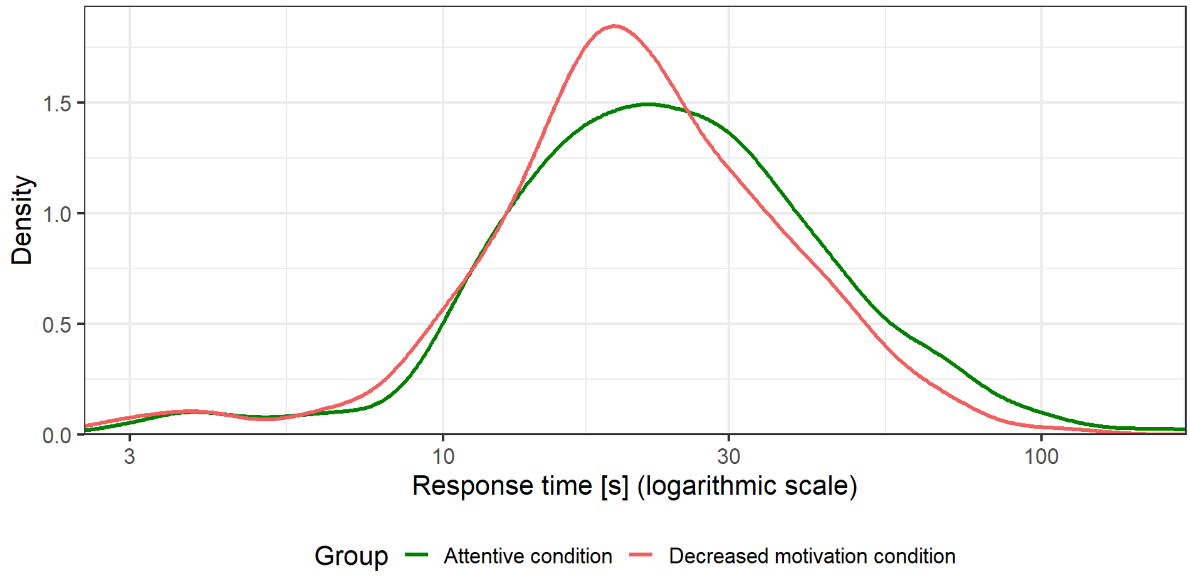


Table A4.11.3

Median and mean log response time to the item M032198

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	28.1	3.14
Decreased motivation	24.3	3.03
ANOVA for mean log(response time) differences: $F(1, 972) = 8.18$, $p = 0.004$		

gmm2014z11 (item set 1)

ICCs and fit statistics

Figure A4.12.1

Item gmm2014z11 item characteristics curves (ICCs)

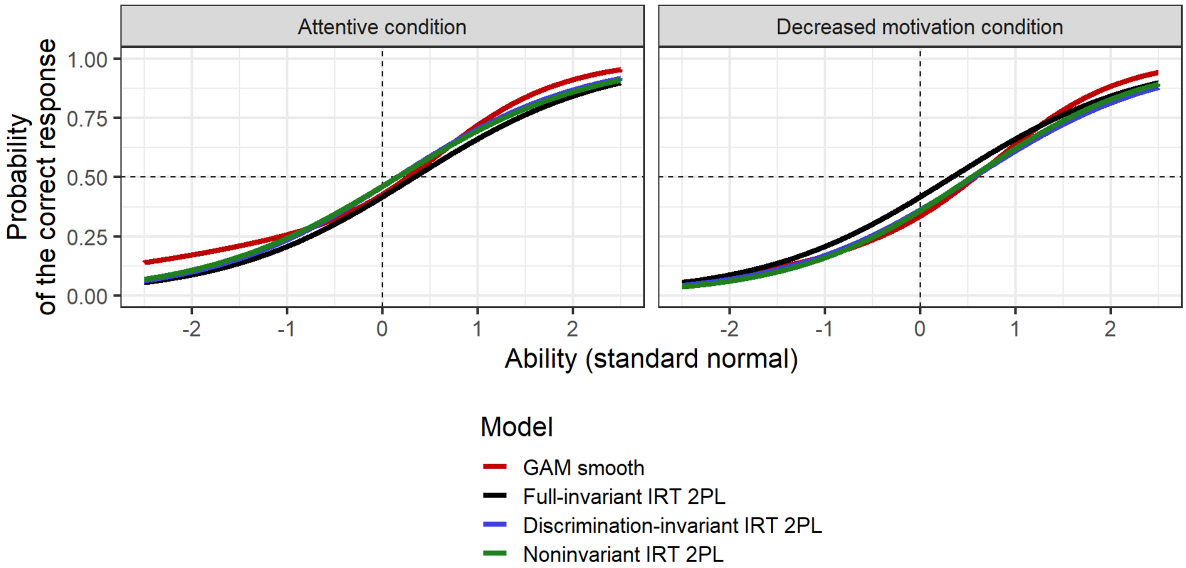


Table A4.12.1

Item gmm2014z11 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.01	0.33		
Attentive condition			0.06	0.003
Decreased motivation condition			0.05	0.021
Discrimination-invariant 2PL model				
All groups	1.02			
Attentive condition		0.15	0.06	0.003
Decreased motivation condition		0.56	0.04	0.035
Noninvariant 2PL model				
Attentive condition	0.98	0.16	0.05	0.004
Decreased motivation condition	1.08	0.54	0.05	0.020

Frequency of answers

Figure A4.12.2

Distribution of answers to the item gmm2014z11

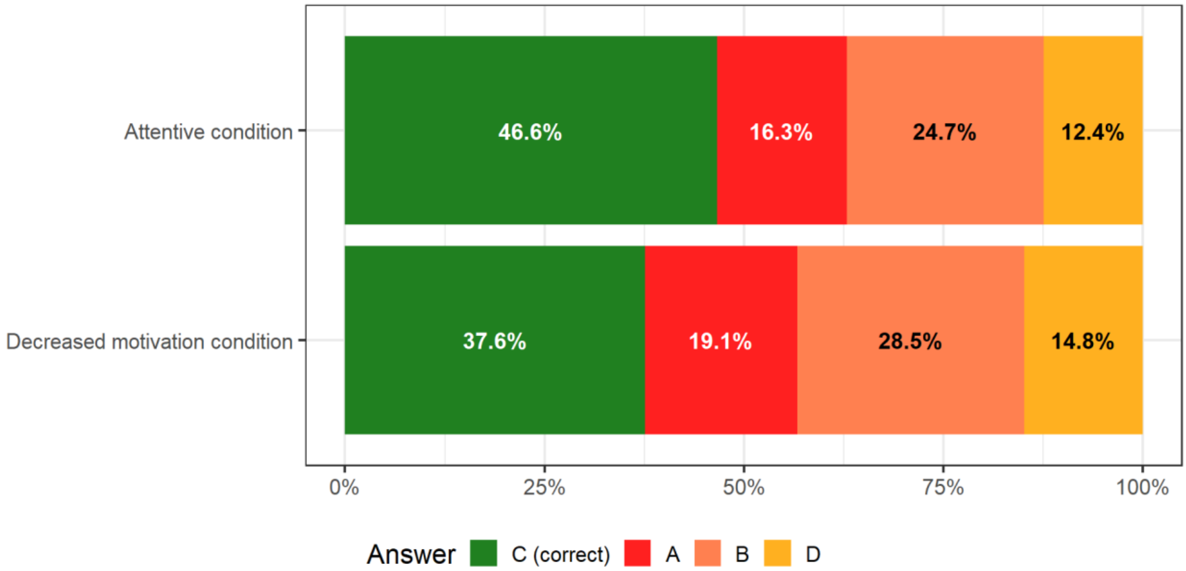


Table A4.12.2

Distribution of answers to the item gmm2014z11

Condition	Answers				Overall
	A	B	C (correct)	D	
Regular (attentive)	87 (16.3%)	132 (24.7%)	249 (46.6%)	66 (12.4%)	534 (100.0%)
Decreased motivation	84 (19.1%)	125 (28.5%)	165 (37.6%)	65 (14.8%)	439 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 8.10$, $p = 0.044$

Answering time

Figure A4.12.3

Distribution of response time to the item gmm2014z11

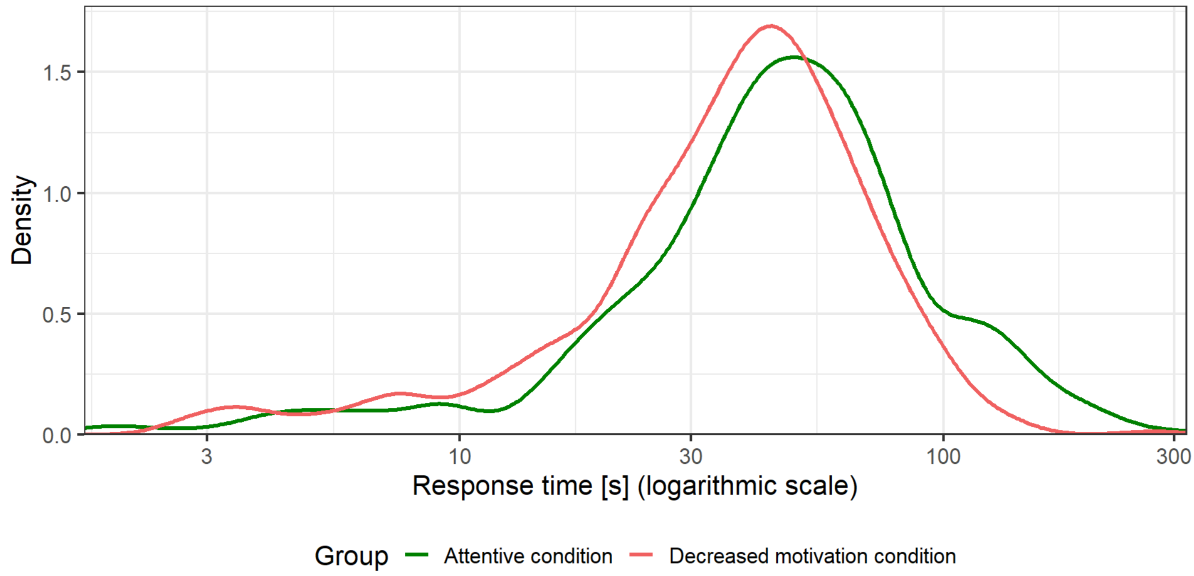


Table A4.12.3

Median and mean log response time to the item gmm2014z11

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	55.3	3.76
Decreased motivation	42.7	3.54
ANOVA for mean log(response time) differences: *F*(1, 972) = 18.71, *p* = 0.000		

M042041 (item set 1)

ICCs and fit statistics

Figure A4.13.1

Item M042041 item characteristics curves (ICCs)

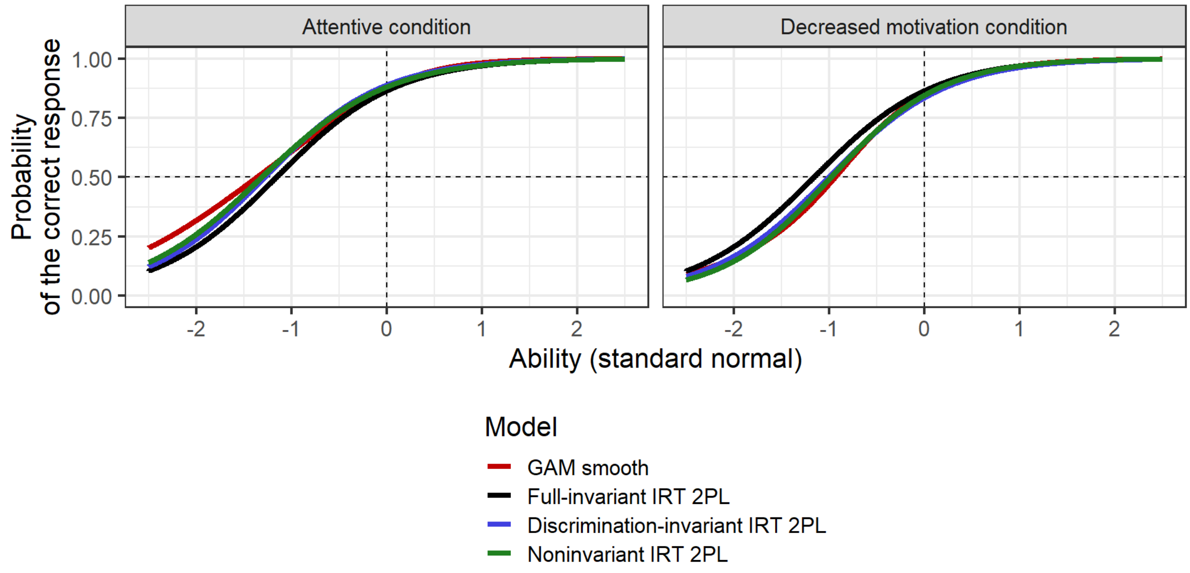


Table A4.13.1

Item M042041 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.60	−1.16		
Attentive condition			0.00	0.514
Decreased motivation condition			0.05	0.019
Discrimination-invariant 2PL model				
All groups	1.62			
Attentive condition		−1.28	0.00	0.500
Decreased motivation condition		−1.00	0.03	0.140
Noninvariant 2PL model				
Attentive condition	1.53	−1.31	0.01	0.409
Decreased motivation condition	1.74	−0.98	0.03	0.141

Frequency of answers

Figure A4.13.2

Distribution of answers to the item M042041

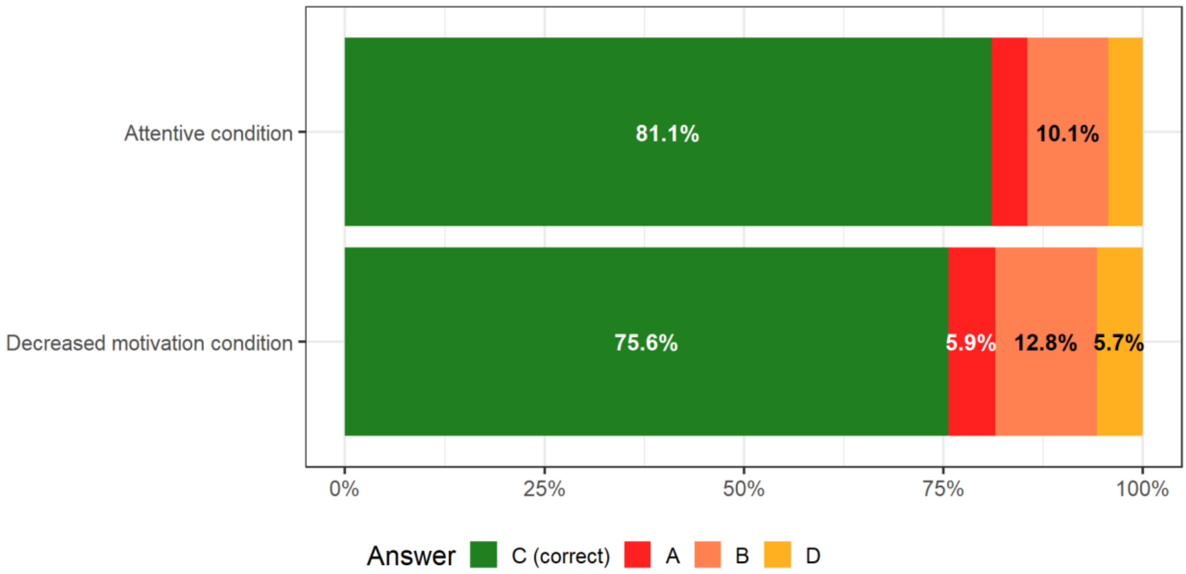


Table A4.13.2

Distribution of answers to the item M042041

Condition	Answers				Overall
	A	B	C (correct)	D	
Regular (attentive)	24 (4.5%)	54 (10.1%)	433 (81.1%)	23 (4.3%)	534 (100.0%)
Decreased motivation	26 (5.9%)	56 (12.8%)	332 (75.6%)	25 (5.7%)	439 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 4.30$, $p = 0.231$

Answering time

Figure A4.13.3

Distribution of response time to the item M042041

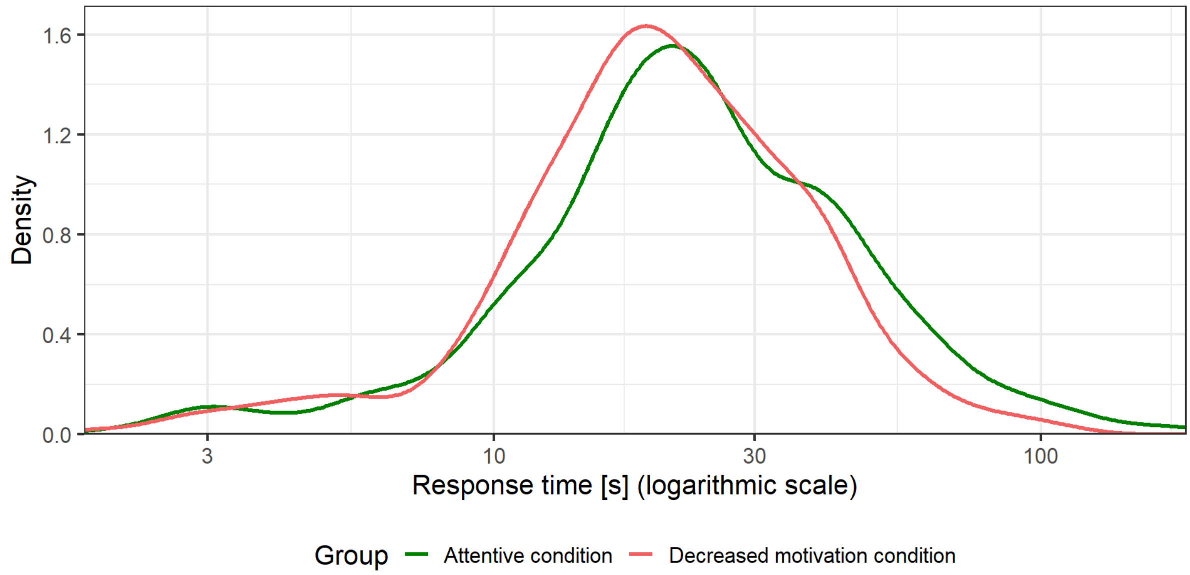


Table A4.13.3

Median and mean log response time to the item M042041

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	28.8	3.12
Decreased motivation	23.7	2.98
ANOVA for mean log(response time) differences: $F(1, 972) = 9.56$, $p = 0.002$		

e82022z7 (item set 1)

ICCs and fit statistics

Figure A4.14.1

Item e82022z7 item characteristics curves (ICCs)

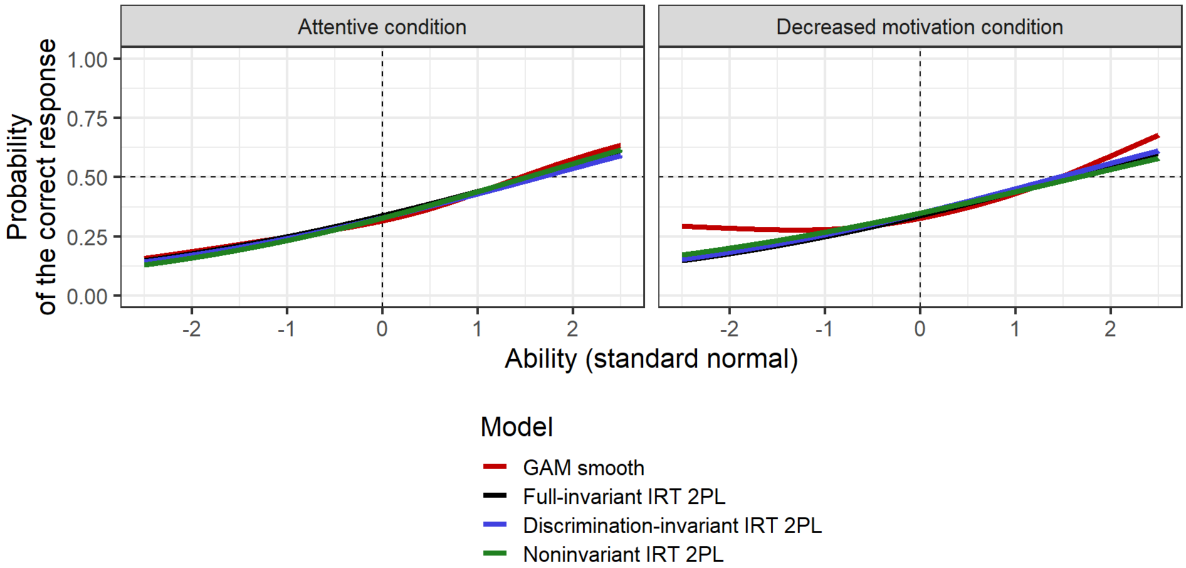


Table A4.14.1

Item e82022z7 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.43	1.56		
Attentive condition			0.03	0.118
Decreased motivation condition			0.03	0.129
Discrimination-invariant 2PL model				
All groups	0.43			
Attentive condition		1.66	0.02	0.268
Decreased motivation condition		1.46	0.04	0.042
Noninvariant 2PL model				
Attentive condition	0.47	1.52	0.02	0.291
Decreased motivation condition	0.38	1.66	0.04	0.029

Frequency of answers

Figure A4.14.2

Distribution of answers to the item e82022z7

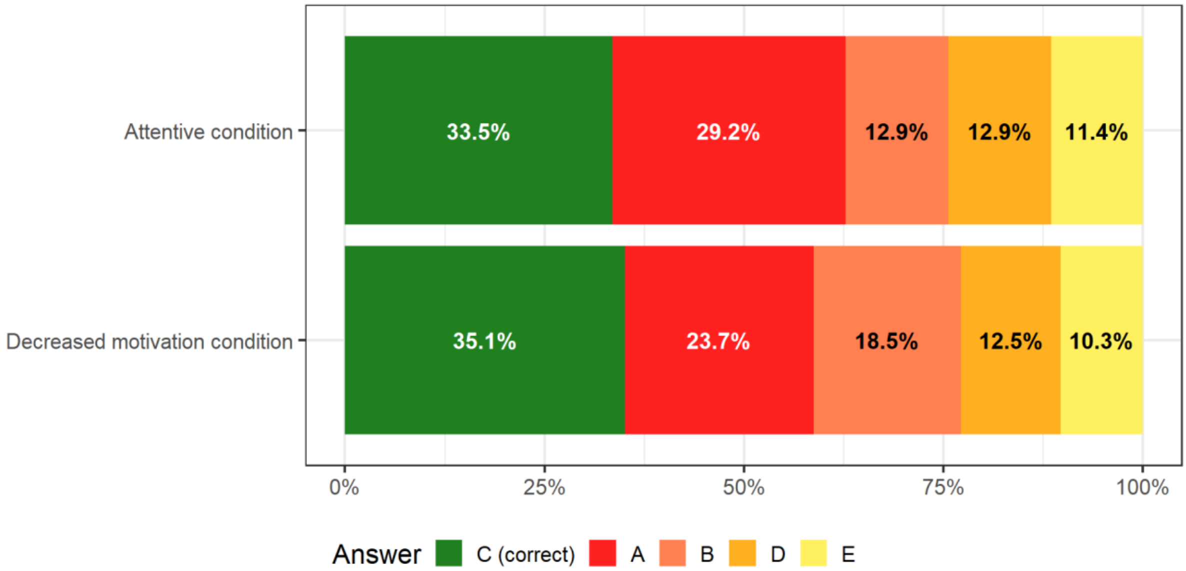


Table A4.14.2

Distribution of answers to the item e82022z7

Condition	Answers					Overall
	A	B	C (correct)	D	E	
Regular (attentive)	156 (29.2%)	69 (12.9%)	179 (33.5%)	69 (12.9%)	61 (11.4%)	534 (100.0%)
Decreased motivation	104 (23.7%)	81 (18.5%)	154 (35.1%)	55 (12.5%)	45 (10.3%)	439 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(4) = 8.03$, $p = 0.090$

Answering time

Figure A4.14.3

Distribution of response time to the item e82022z7

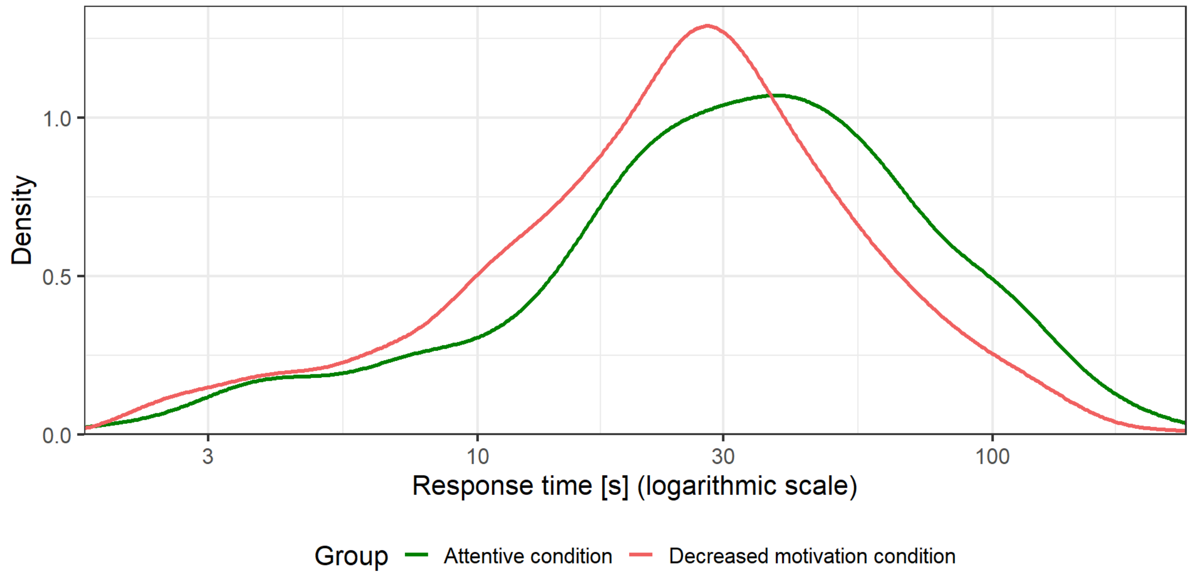


Table A4.14.3

Median and mean log response time to the item e82022z7

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	42.8	3.40
Decreased motivation	32.2	3.15
ANOVA for mean log(response time) differences: *F*(1, 972) = 20.08, *p* = 0.000		

M032166 (item set 1)

ICCs and fit statistics

Figure A4.15.1

Item M032166 item characteristics curves (ICCs)

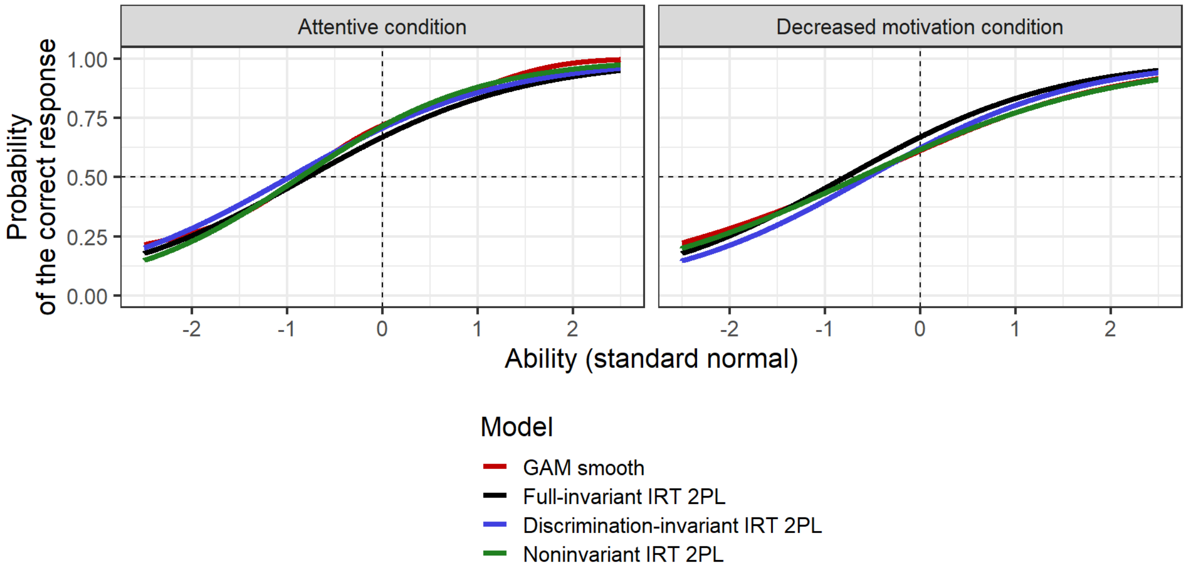


Table A4.15.1

Item M032166 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.90	−0.79		
Attentive condition			0.03	0.118
Decreased motivation condition			0.04	0.031
Discrimination-invariant 2PL model				
All groups	0.90			
Attentive condition		−0.98	0.04	0.047
Decreased motivation condition		−0.55	0.01	0.398
Noninvariant 2PL model				
Attentive condition	1.07	−0.87	0.04	0.075
Decreased motivation condition	0.74	−0.64	0.04	0.036

Frequency of answers

Figure A4.15.2

Distribution of answers to the item M032166

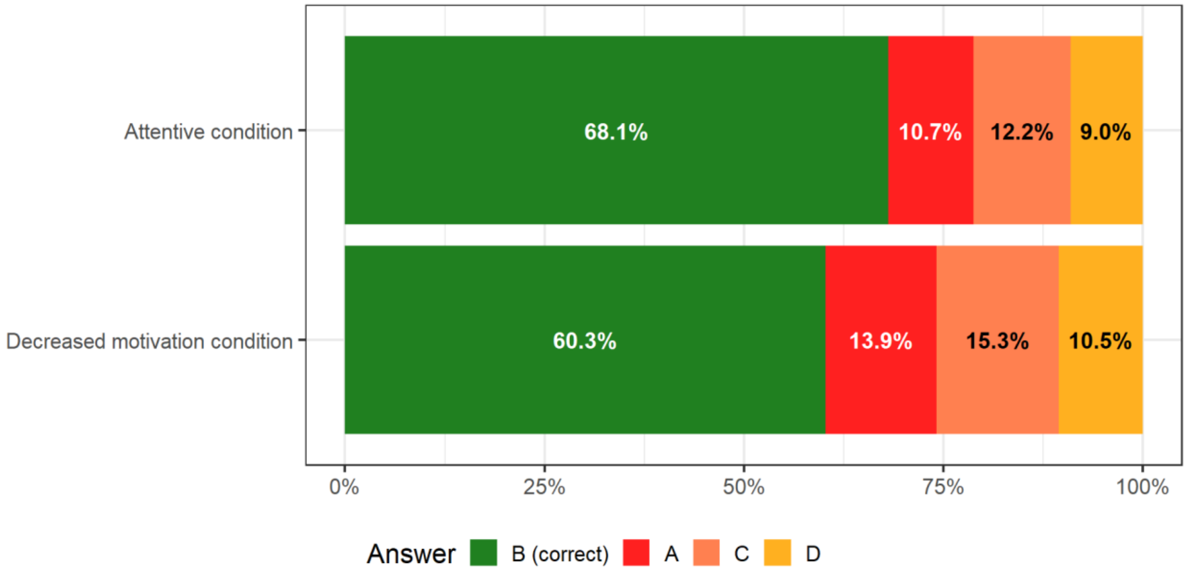


Table A4.15.2

Distribution of answers to the item M032166

Condition	Answers				Overall
	A	**B (correct)**	**C**	**D**	
Regular (attentive)	57 (10.7%)	363 (68.1%)	65 (12.2%)	48 (9.0%)	533 (100.0%)
Decreased motivation	61 (13.9%)	264 (60.3%)	67 (15.3%)	46 (10.5%)	438 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 6.61$, $p = 0.085$

Answering time

Figure A4.15.3

Distribution of response time to the item M032166

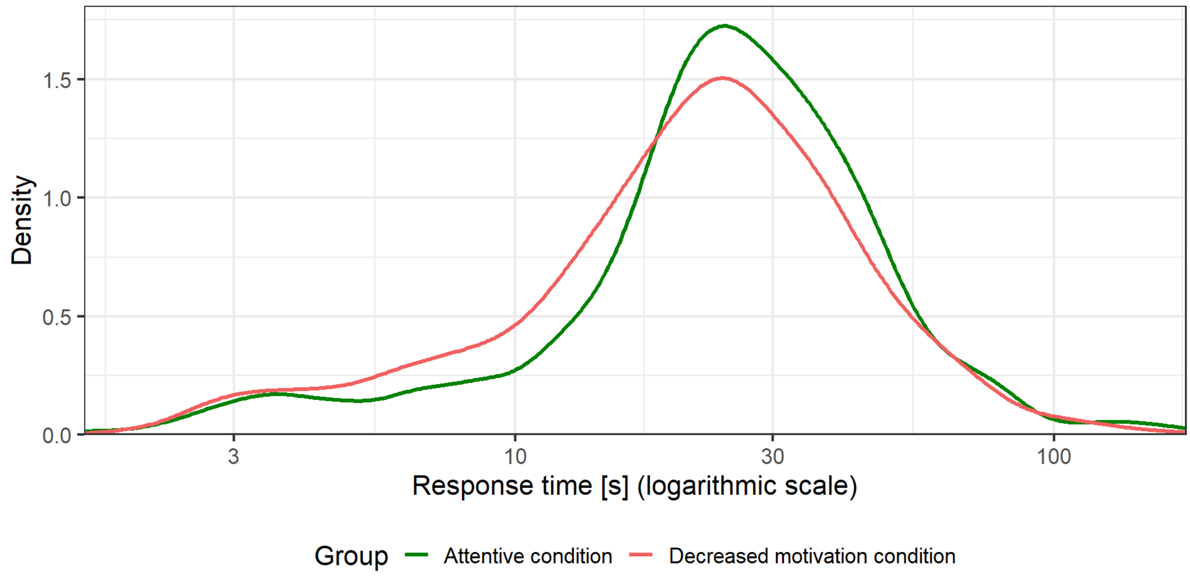


Table A4.15.3

Median and mean log response time to the item M032166

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	29.1	3.15
Decreased motivation	26.0	3.02
ANOVA for mean log(response time) differences: $F(1, 972) = 7.78$, $p = 0.005$		

M042032 (item set 2)

ICCs and fit statistics

Figure A4.16.1

Item M042032 item characteristics curves (ICCs)

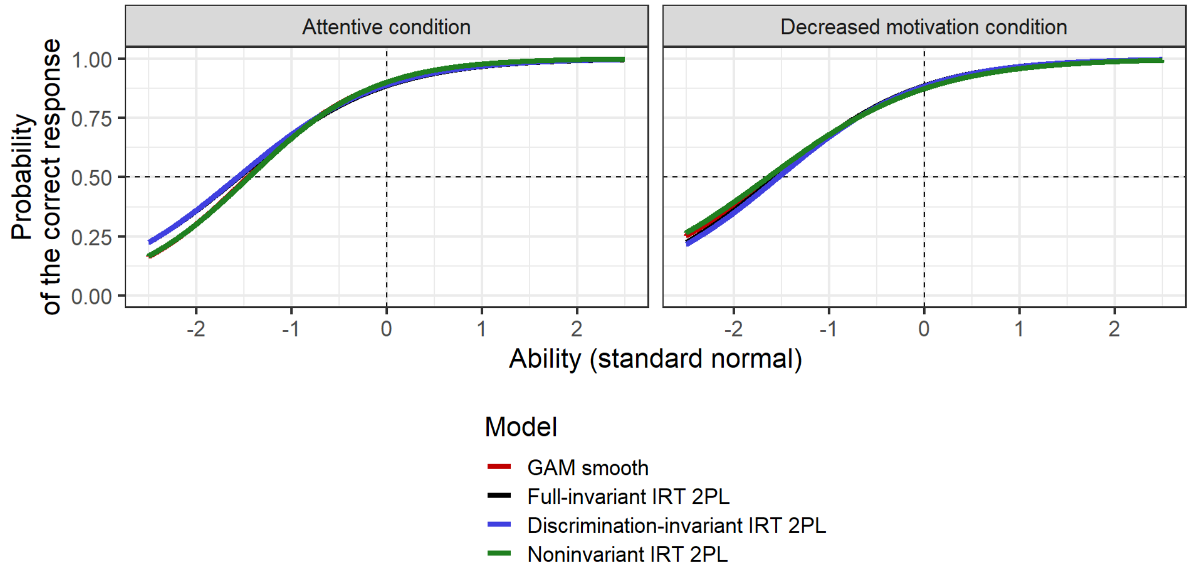


Table A4.16.1

Item M042032 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.32	−1.56		
Attentive condition			0.02	0.272
Decreased motivation condition			0.00	0.499
Discrimination-invariant 2PL model				
All groups	1.33			
Attentive condition		−1.57	0.03	0.189
Decreased motivation condition		−1.53	0.01	0.404
Noninvariant 2PL model				
Attentive condition	1.52	−1.45	0.00	0.550
Decreased motivation condition	1.18	−1.64	0.03	0.117

Frequency of answers

Figure A4.16.2

Distribution of answers to the item M042032

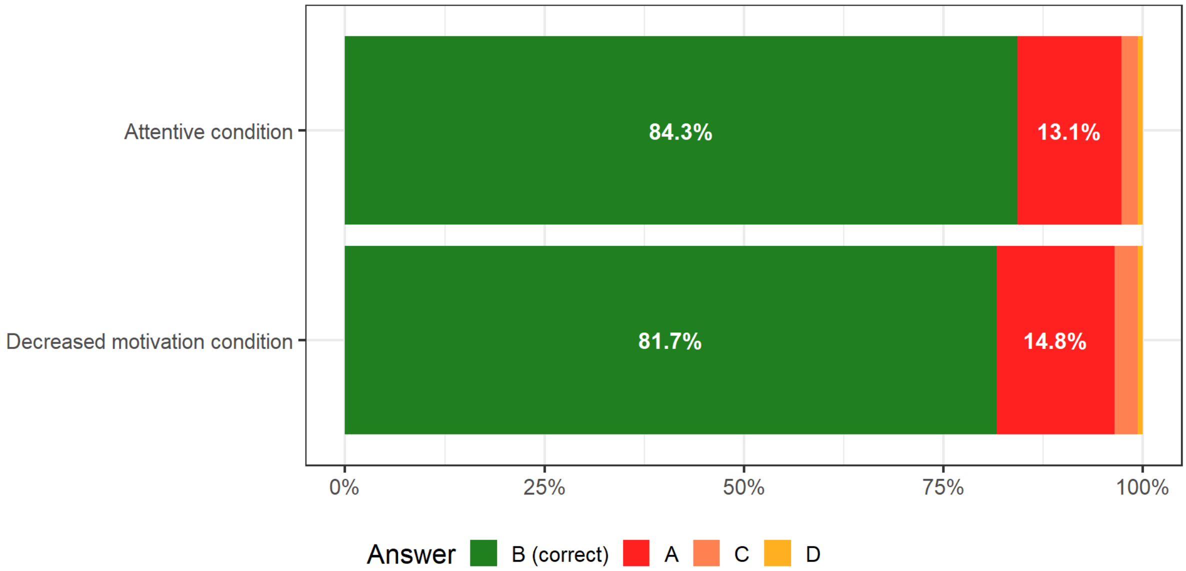


Table A4.16.2

Distribution of answers to the item M042032

Condition	Answers				Overall
	A	**B (correct)**	**C**	**D**	
Regular (attentive)	65 (13.1%)	418 (84.3%)	10 (2.0%)	3 (0.6%)	496 (100.0%)
Decreased motivation	72 (14.8%)	397 (81.7%)	14 (2.9%)	3 (0.6%)	486 (100.0%)

Tests of independence: p-value was simulated because of the low frequencies in some cells of the joint distribution, p = 0.691

Answering time

Figure A4.16.3

Distribution of response time to the item M042032

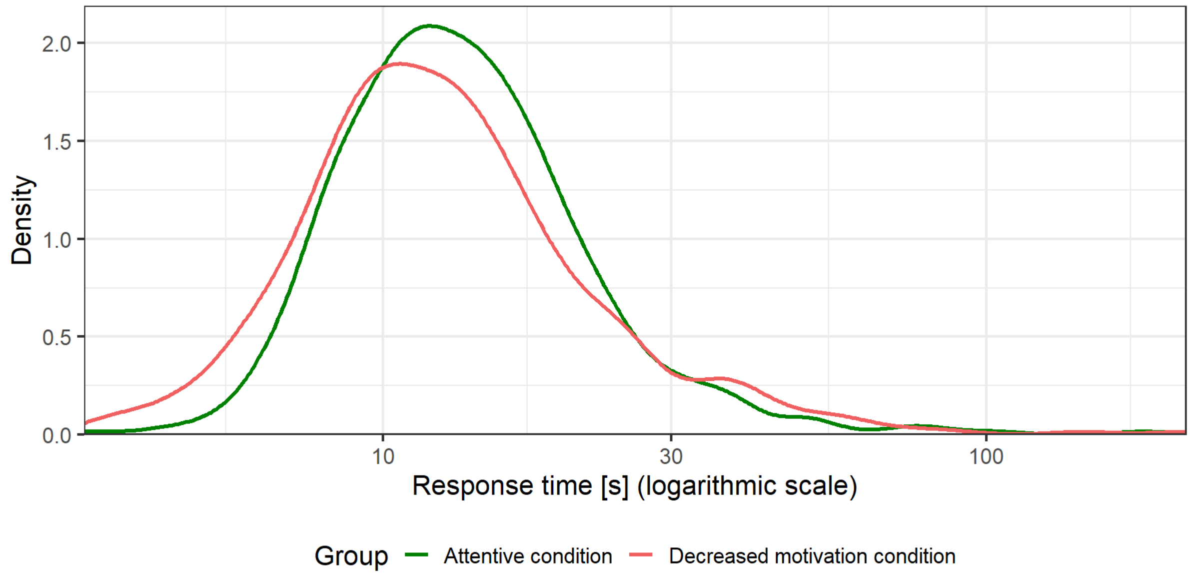


Table A4.16.3

Median and mean log response time to the item M042032

Condition	**Median response time [s]**	**Mean log(response time)**
Regular (attentive)	15.6	2.61
Decreased motivation	15.4	2.54
ANOVA for mean log(response time) differences: $F(1, 980) = 4.22$, $p = 0.040$		

M052302 (item set 2)

ICCs and fit statistics

Figure A4.17.1

Item M052302 item characteristics curves (ICCs)

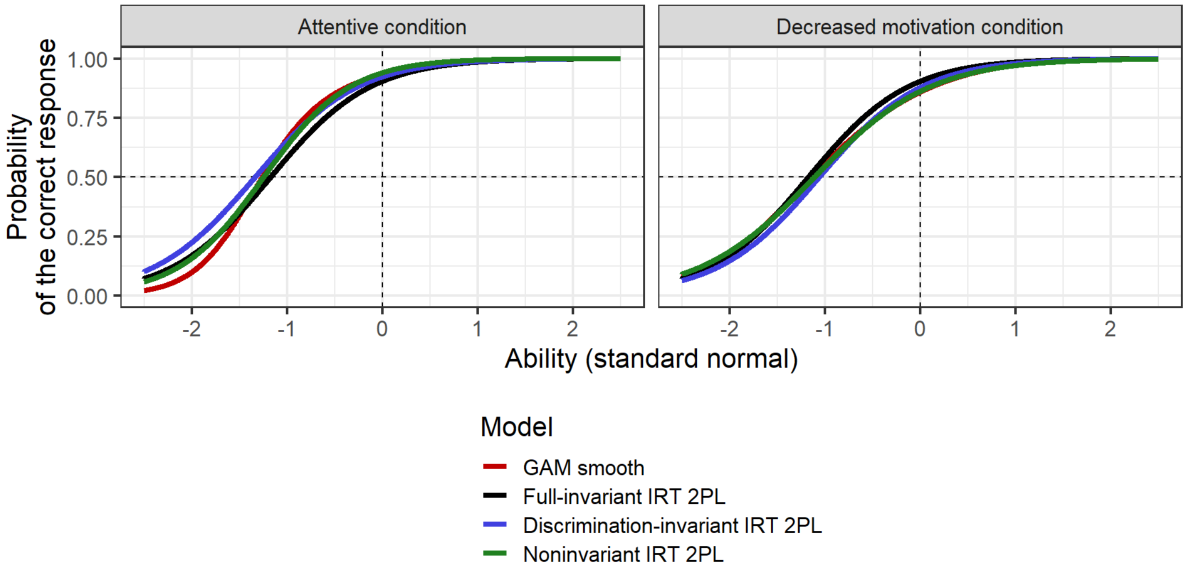


Table A4.17.1

Item M052302 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.92	−1.17		
Attentive condition			0.05	0.021
Decreased motivation condition			0.04	0.040
Discrimination-invariant 2PL model				
All groups	1.87			
Attentive condition		−1.34	0.05	0.011
Decreased motivation condition		−1.06	0.02	0.248
Noninvariant 2PL model				
Attentive condition	2.22	−1.25	0.04	0.062
Decreased motivation condition	1.65	−1.11	0.02	0.310

Frequency of answers

Figure A4.17.2

Distribution of answers to the item M052302

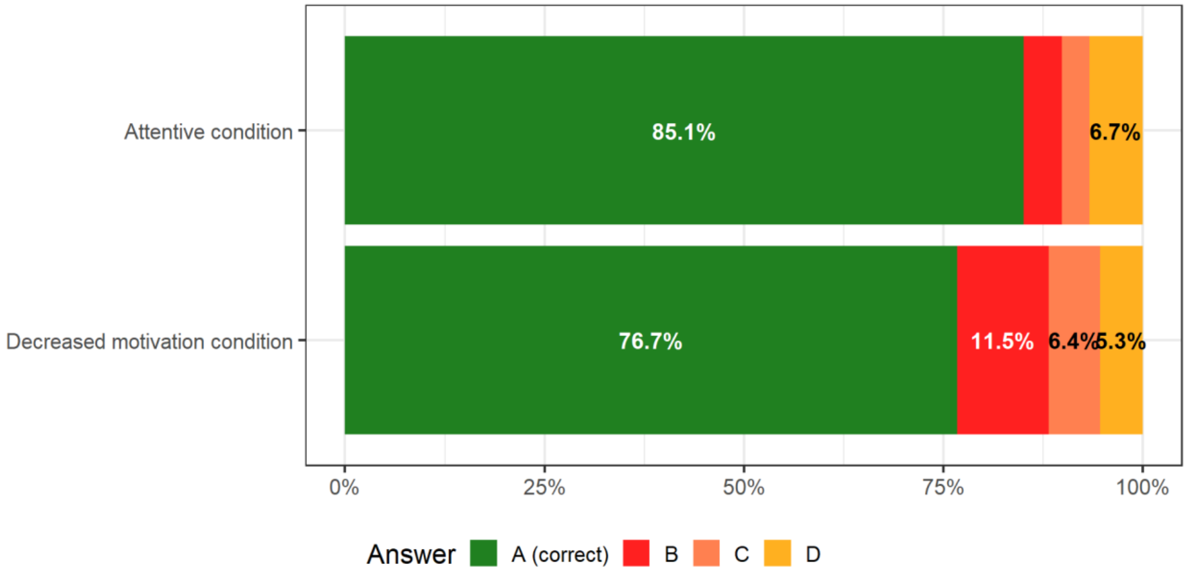


Table A4.17.2

Distribution of answers to the item M052302

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	422 (85.1%)	24 (4.8%)	17 (3.4%)	33 (6.7%)	496 (100.0%)
Decreased motivation	373 (76.7%)	56 (11.5%)	31 (6.4%)	26 (5.3%)	486 (100.0%)
Pearson Chi-squared tests of independence: $\chi^2(3) = 20.63$, $p = 0.000$					

Answering time

Figure A4.17.3

Distribution of response time to the item M052302

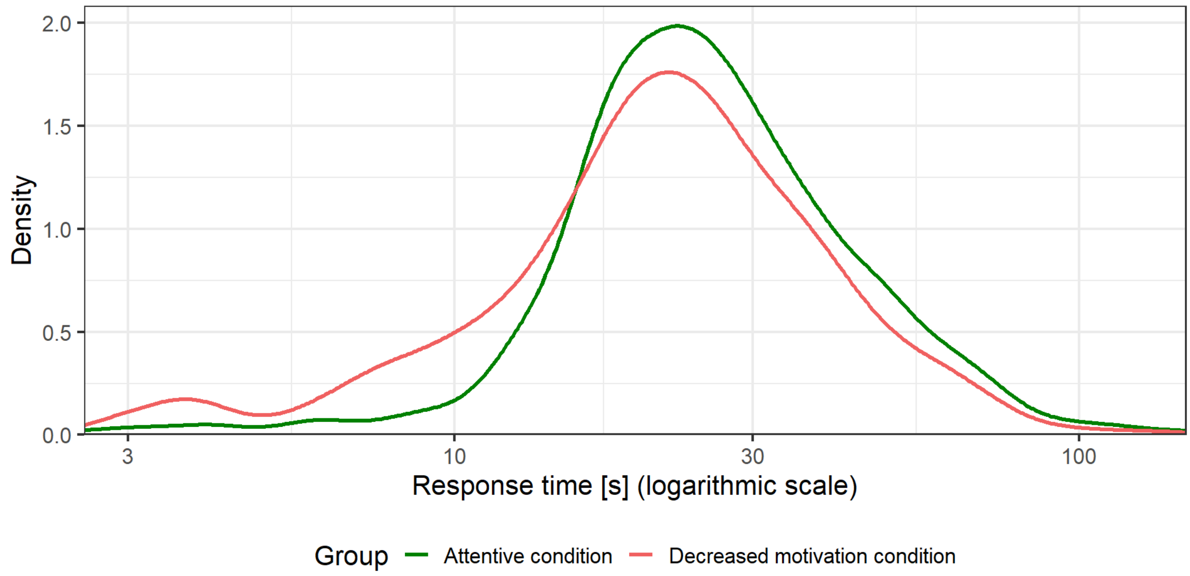


Table A4.17.3

Median and mean log response time to the item M052302

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	29.0	3.23
Decreased motivation	25.0	3.03
ANOVA for mean log(response time) differences: $F(1, 980) = 26.18$, $p = 0.000$		

gmm2016z20 (item set 2)

ICCs and fit statistics

Figure A4.18.1

Item gmm2016z20 item characteristics curves (ICCs)

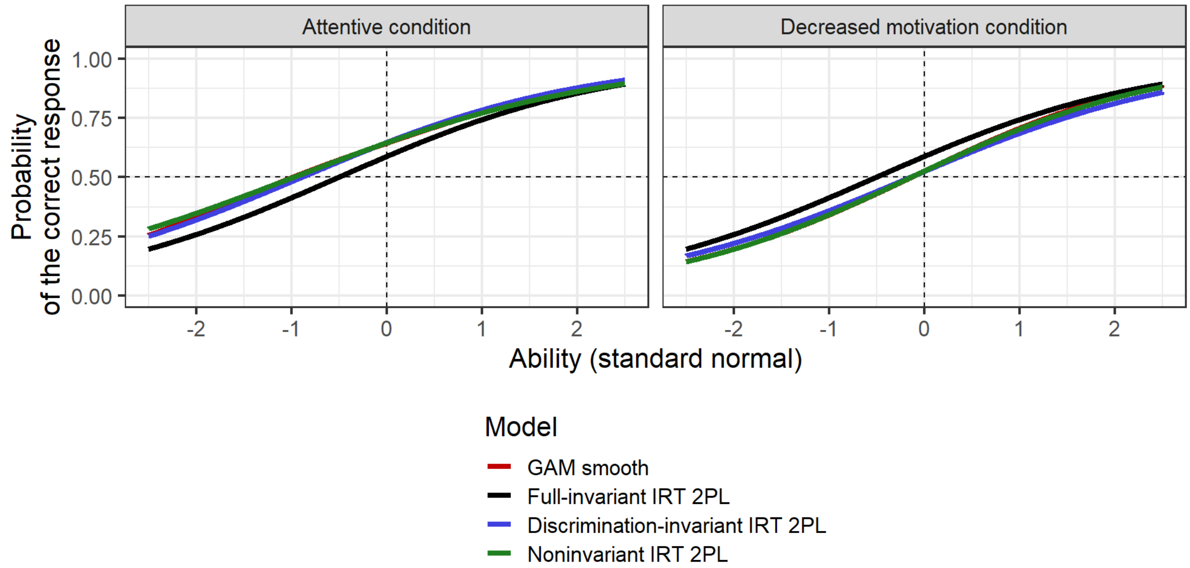


Table A4.18.1

Item gmm2016z20 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.71	−0.50		
Attentive condition			0.05	0.027
Decreased motivation condition			0.04	0.036
Discrimination-invariant 2PL model				
All groups	0.68			
Attentive condition		−0.89	0.01	0.372
Decreased motivation condition		−0.14	0.03	0.105
Noninvariant 2PL model				
Attentive condition	0.61	−0.97	0.02	0.219
Decreased motivation condition	0.76	−0.14	0.04	0.049

Frequency of answers

Figure A4.18.2

Distribution of answers to the item gmm2016z20

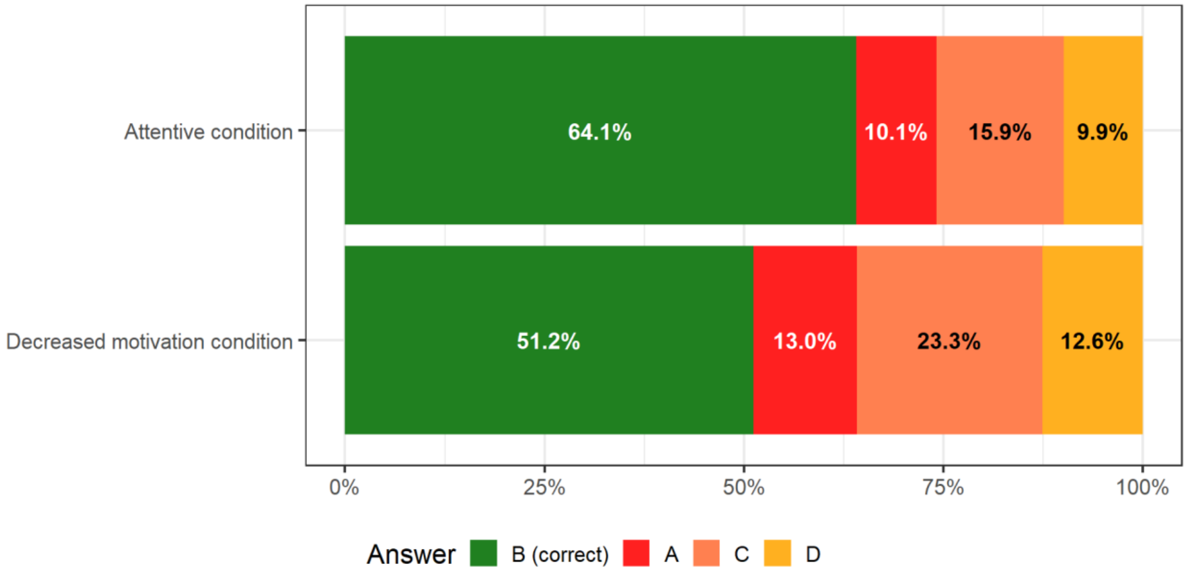


Table A4.18.2

Distribution of answers to the item gmm2016z20

Condition	**Answers**				**Overall**
	A	**B (correct)**	**C**	**D**	
Regular (attentive)	50 (10.1%)	318 (64.1%)	79 (15.9%)	49 (9.9%)	496 (100.0%)
Decreased motivation	63 (13.0%)	249 (51.2%)	113 (23.3%)	61 (12.6%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 17.12$, $p = 0.001$

Answering time

Figure A4.18.3

Distribution of response time to the item gmm2016z20

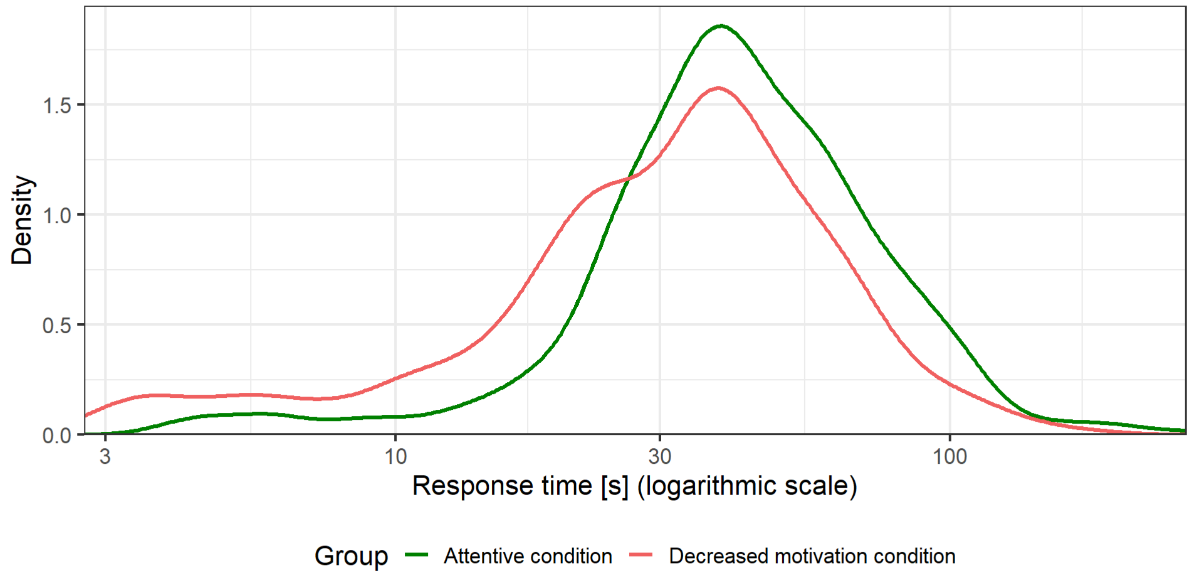


Table A4.18.3

Median and mean log response time to the item gmm2016z20

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	47.5	3.69
Decreased motivation	36.7	3.36
ANOVA for mean log(response time) differences: $F(1, 980) = 55.05$, $p = 0.000$		

M032738 (item set 2)

ICCs and fit statistics

Figure A4.19.1

Item M032738 item characteristics curves (ICCs)

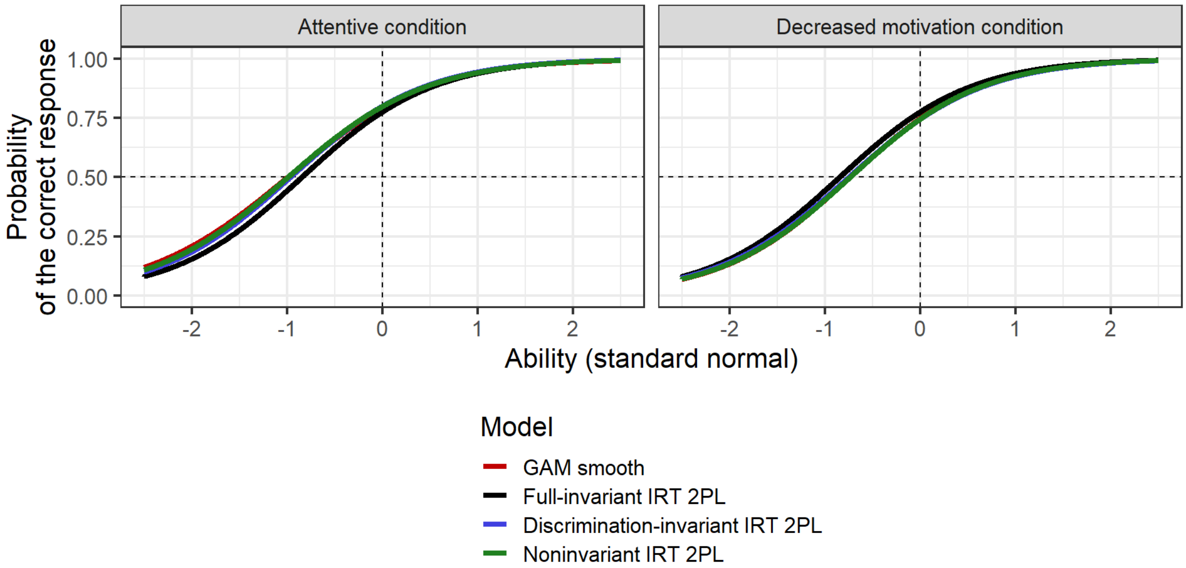


Table A4.19.1

Item M032738 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.47	−0.84		
Attentive condition			0.03	0.123
Decreased motivation condition			0.03	0.109
Discrimination-invariant 2PL model				
All groups	1.43			
Attentive condition		−0.97	0.04	0.079
Decreased motivation condition		−0.75	0.02	0.289
Noninvariant 2PL model				
Attentive condition	1.39	−0.99	0.04	0.062
Decreased motivation condition	1.46	−0.73	0.04	0.076

Frequency of answers

Figure A4.19.2

Distribution of answers to the item M032738

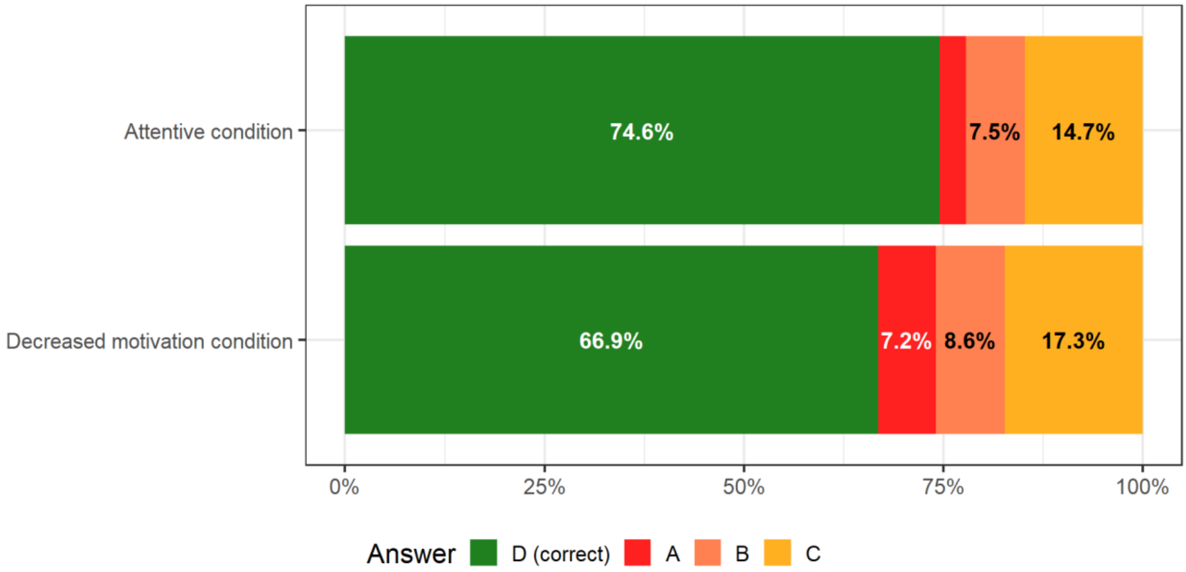


Table A4.19.2

Distribution of answers to the item M032738

Condition	Answers				Overall
	A	**B**	**C**	**D (correct)**	
Regular (attentive)	16 (3.2%)	37 (7.5%)	73 (14.7%)	370 (74.6%)	496 (100.0%)
Decreased motivation	35 (7.2%)	42 (8.6%)	84 (17.3%)	325 (66.9%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 10.98$, $p = 0.012$

Answering time

Figure A4.19.3

Distribution of response time to the item M032738

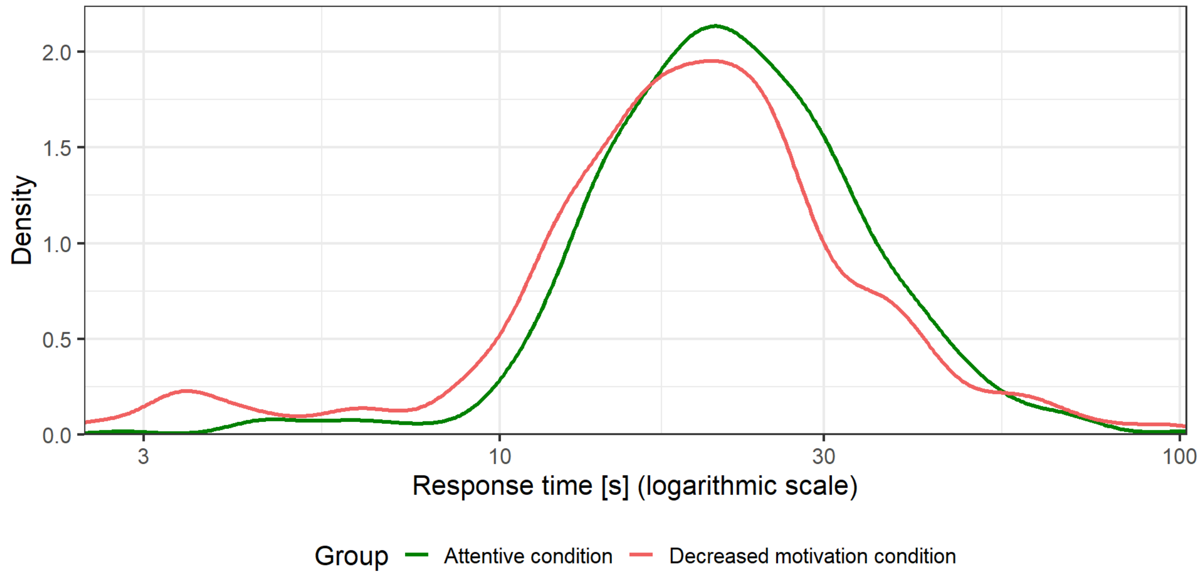


Table A4.19.3

Median and mean log response time to the item M032738

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	23.9	3.07
Decreased motivation	21.8	2.92
ANOVA for mean log(response time) differences: $F(1, 980) = 20.24$, $p = 0.000$		

M032477 (item set 2)

ICCs and fit statistics

Figure A4.20.1

Item M032477 item characteristics curves (ICCs)

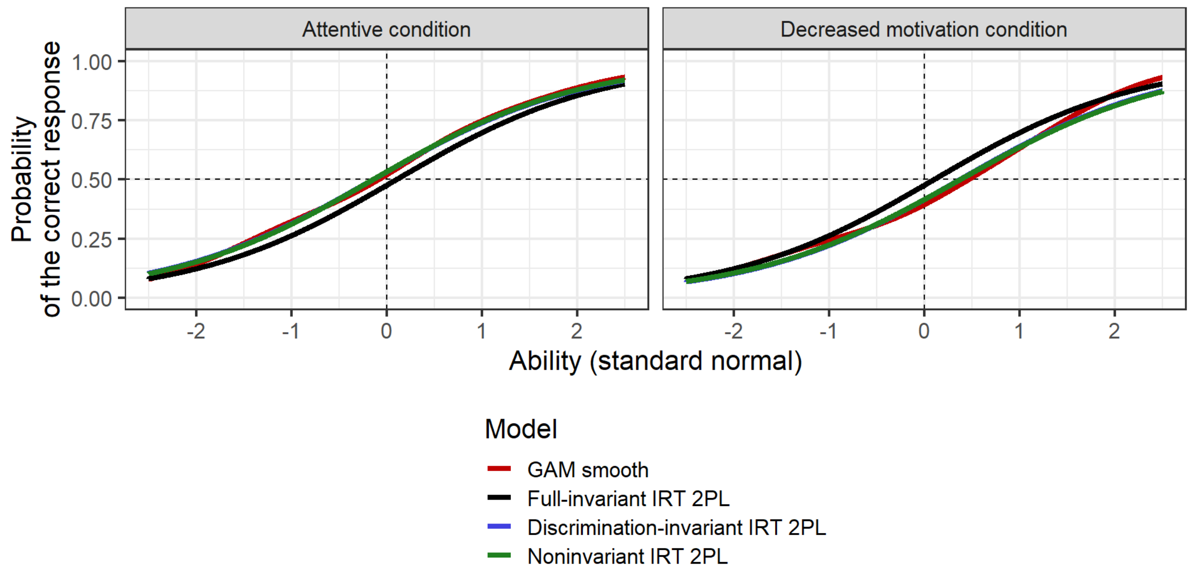


Table A4.20.1

Item M032477 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.93	0.10		
Attentive condition			0.03	0.152
Decreased motivation condition			0.04	0.052
Discrimination-invariant 2PL model				
All groups	0.91			
Attentive condition		−0.14	0.04	0.037
Decreased motivation condition		0.37	0.03	0.184
Noninvariant 2PL model				
Attentive condition	0.92	−0.14	0.04	0.047
Decreased motivation condition	0.90	0.38	0.02	0.223

Frequency of answers

Figure A4.20.2

Distribution of answers to the item M032477

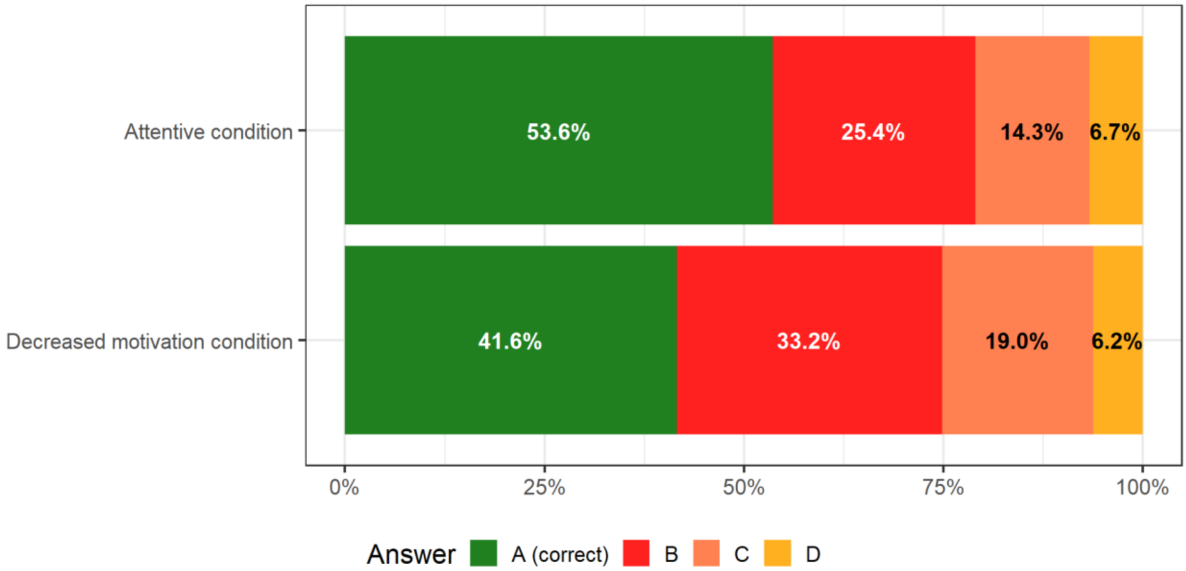


Table A4.20.2

Distribution of answers to the item M032477

Condition	Answers				Overall
	A (correct)	**B**	**C**	**D**	
Regular (attentive)	266 (53.6%)	126 (25.4%)	71 (14.3%)	33 (6.7%)	496 (100.0%)
Decreased motivation	202 (41.6%)	161 (33.2%)	92 (19.0%)	30 (6.2%)	485 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 15.75$, $p = 0.001$

Answering time

Figure A4.20.3

Distribution of response time to the item M032477

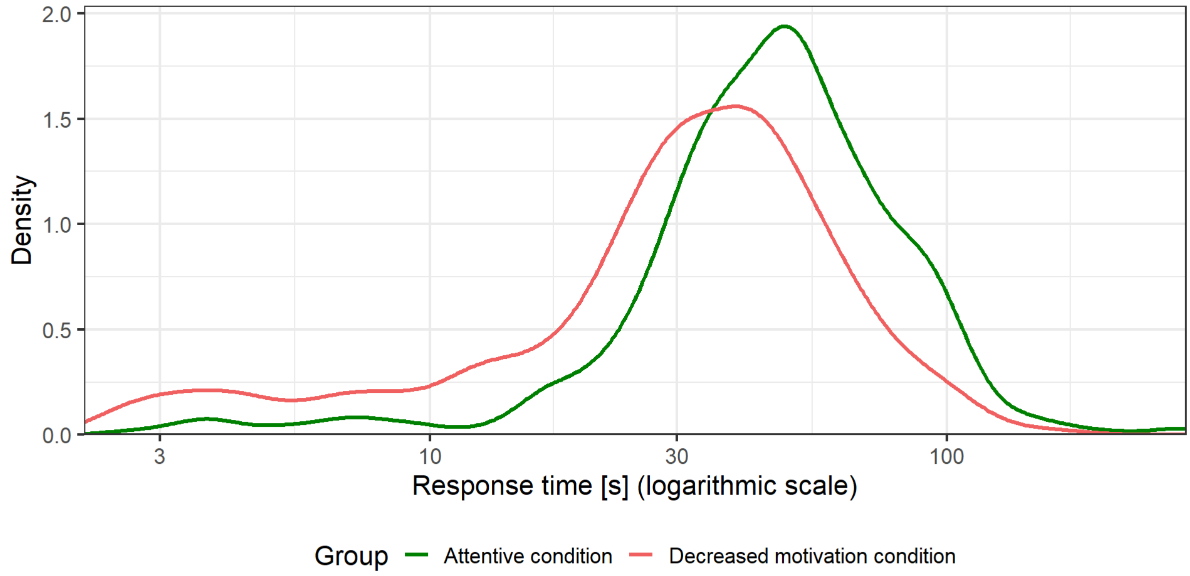


Table A4.20.3

Median and mean log response time to the item M032477

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	52.0	3.79
Decreased motivation	36.3	3.33
ANOVA for mean log(response time) differences: *F*(1, 980) = 94.55, *p* = 0.000		

gmm2018z07 (item set 2)

ICCs and fit statistics

Figure A4.21.1

Item gmm2018z07 item characteristics curves (ICCs)

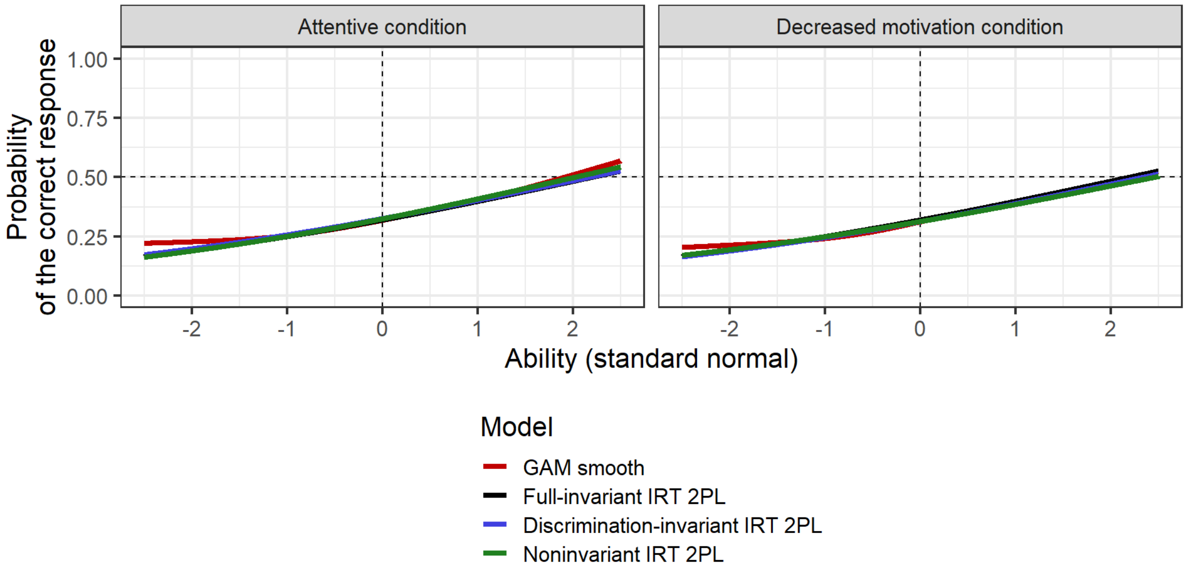


Table A4.21.1

Item gmm2018z07 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.34	2.20		
Attentive condition			0.03	0.176
Decreased motivation condition			0.03	0.184
Discrimination-invariant 2PL model				
All groups	0.34			
Attentive condition		2.17	0.01	0.388
Decreased motivation condition		2.35	0.03	0.168
Noninvariant 2PL model				
Attentive condition	0.36	2.02	0.03	0.110
Decreased motivation condition	0.32	2.47	0.03	0.125

Frequency of answers

Figure A4.21.2

Distribution of answers to the item gmm2018z07

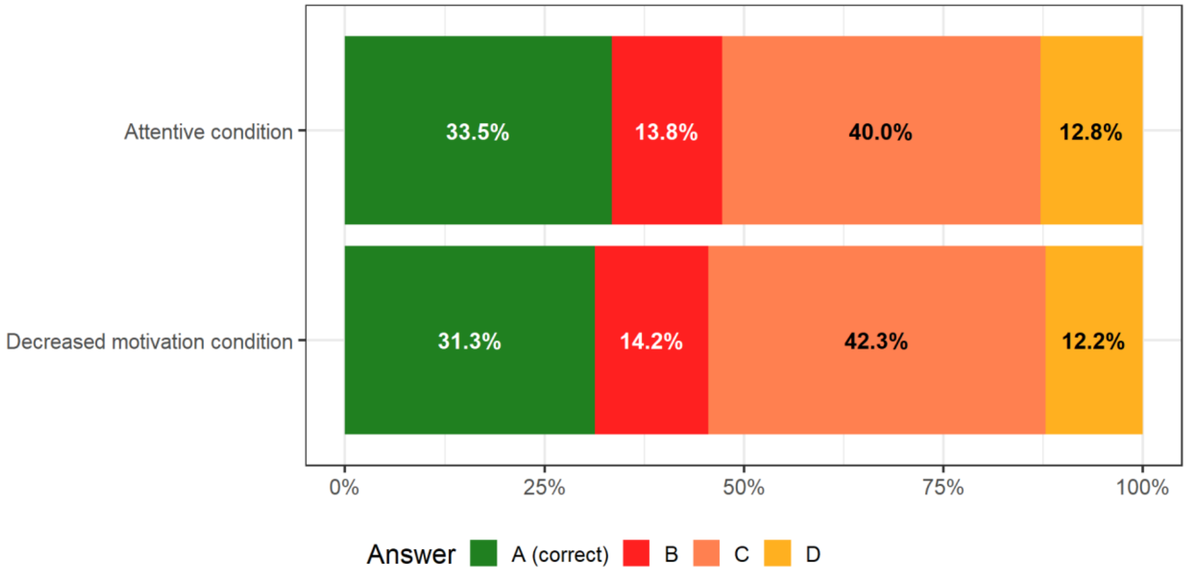


Table A4.21.2

Distribution of answers to the item gmm2018z07

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	165 (33.5%)	68 (13.8%)	197 (40.0%)	63 (12.8%)	493 (100.0%)
Decreased motivation	152 (31.3%)	69 (14.2%)	205 (42.3%)	59 (12.2%)	485 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 0.77$, $p = 0.858$

Answering time

Figure A4.21.3

Distribution of response time to the item gmm2018z07

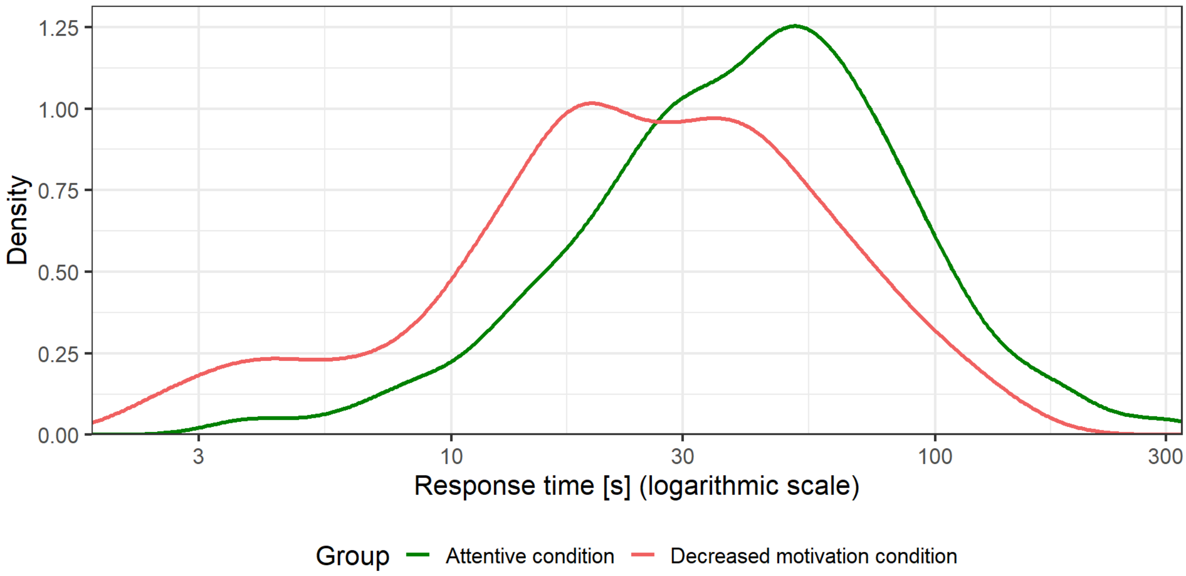


Table A4.21.3

Median and mean log response time to the item gmm2018z07

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	52.8	3.69
Decreased motivation	33.4	3.16
ANOVA for mean log(response time) differences: $F(1, 980) = 99.60$, $p = 0.000$		

M032704 (item set 2)

ICCs and fit statistics

Figure A4.22.1

Item M032704 item characteristics curves (ICCs)

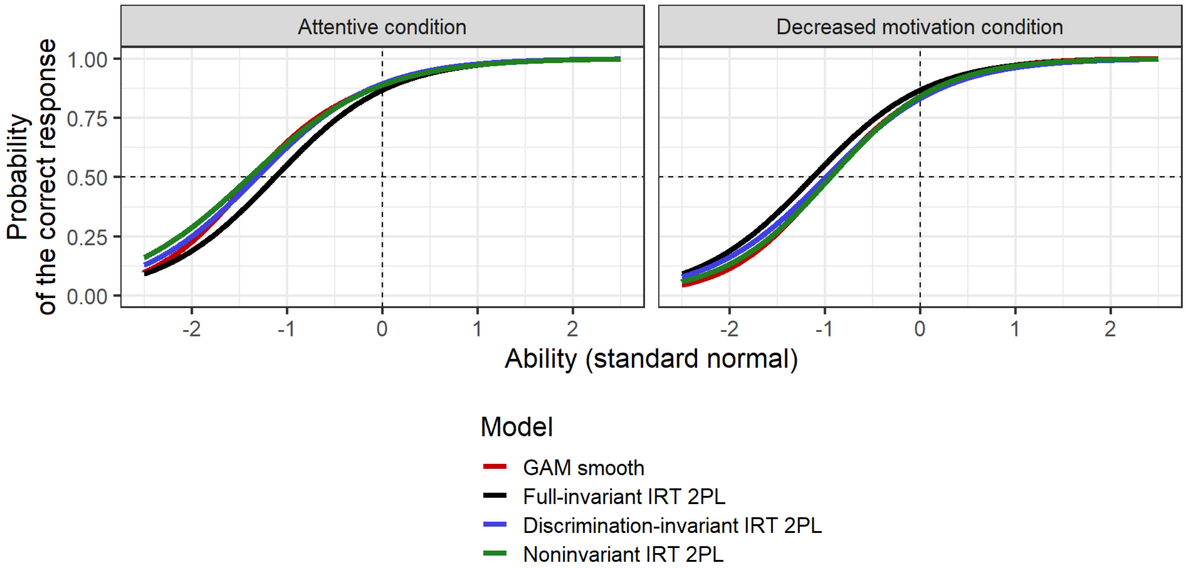


Table A4.22.1

Item M032704 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.67	−1.13		
Attentive condition			0.00	0.880
Decreased motivation condition			0.03	0.098
Discrimination-invariant 2PL model				
All groups	1.62			
Attentive condition		−1.32	0.00	0.798
Decreased motivation condition		−0.99	0.02	0.223
Noninvariant 2PL model				
Attentive condition	1.49	−1.39	0.00	0.757
Decreased motivation condition	1.77	−0.94	0.00	0.507

Frequency of answers

Figure A4.22.2

Distribution of answers to the item M032704

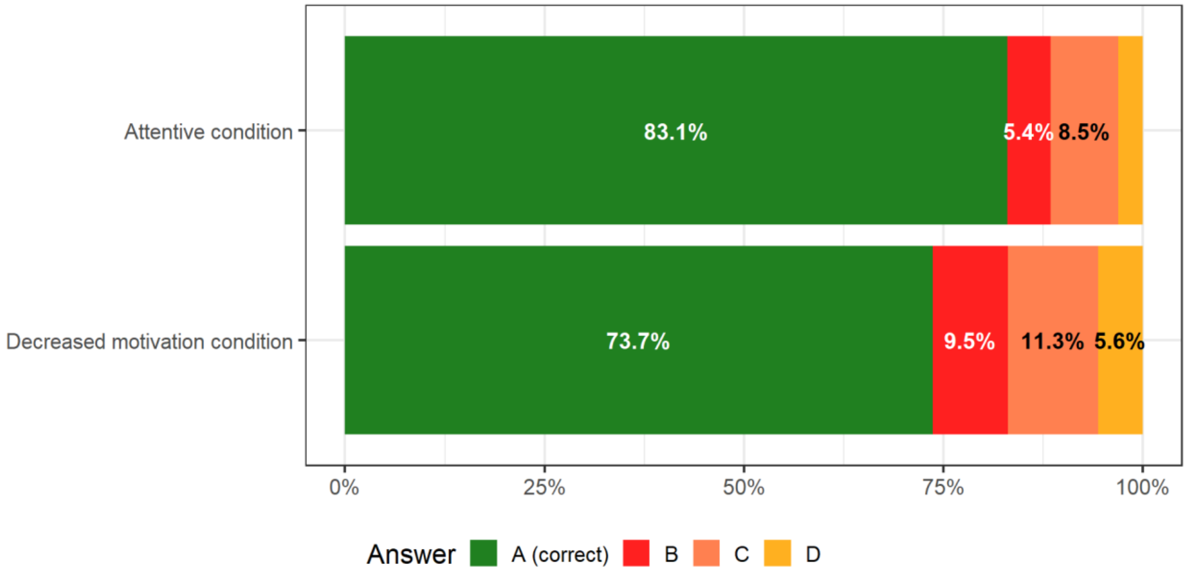


Table A4.22.2

Distribution of answers to the item M032704

Condition	Answers				Overall
	A (correct)	**B**	**C**	**D**	
Regular (attentive)	412 (83.1%)	27 (5.4%)	42 (8.5%)	15 (3.0%)	4963 (100.0%)
Decreased motivation	358 (73.7%)	46 (9.5%)	55 (11.3%)	27 (5.6%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 13.80$, $p = 0.003$

Answering time

Figure A4.22.3

Distribution of response time to the item M032704

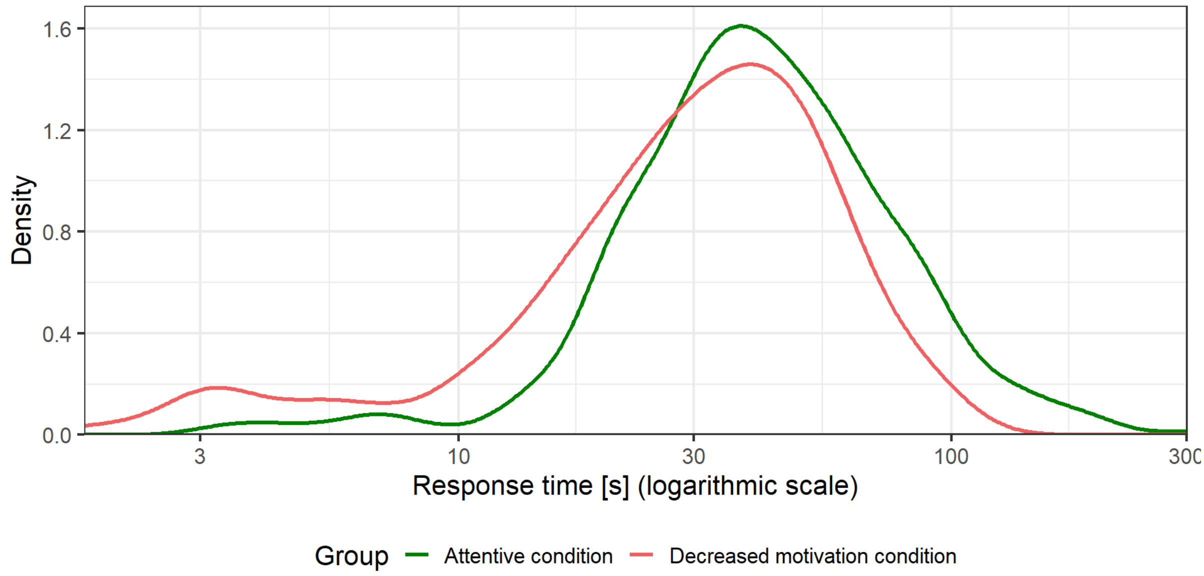


Table A4.22.3

Median and mean log response time to the item M032704

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	48.5	3.69
Decreased motivation	34.5	3.30
ANOVA for mean log(response time) differences: $F(1, 980) = 70.30$, $p = 0.000$		

M032116 (item set 2)

ICCs and fit statistics

Figure A4.23.1

Item M032116 item characteristics curves (ICCs)

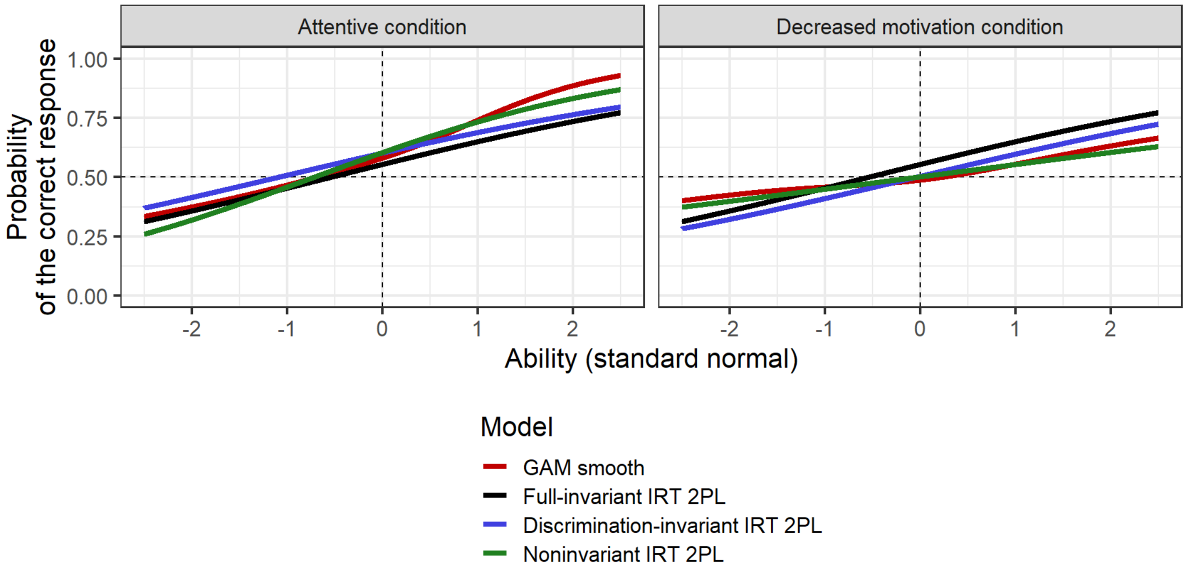


Table A4.23.1

Item M032116 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.40	−0.54		
Attentive condition			0.06	0.004
Decreased motivation condition			0.03	0.153
Discrimination-invariant 2PL model				
All groups	0.38			
Attentive condition		−1.08	0.05	0.009
Decreased motivation condition		−0.03	0.01	0.371
Noninvariant 2PL model				
Attentive condition	0.59	−0.71	0.02	0.249
Decreased motivation condition	0.21	−0.01	0.01	0.367

Frequency of answers

Figure A4.23.2

Distribution of answers to the item M032116

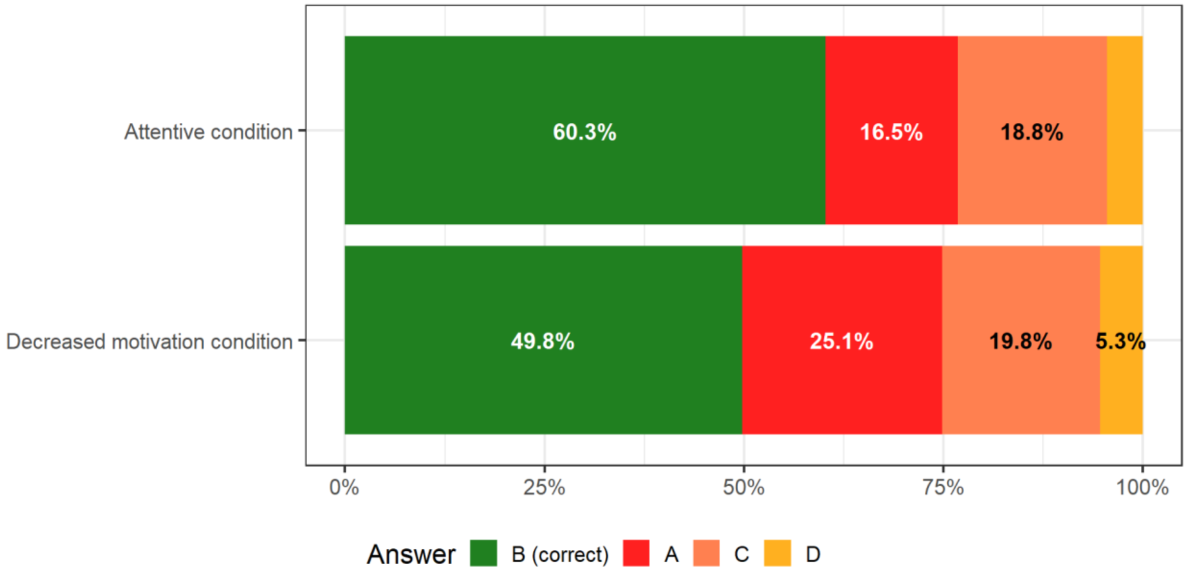


Table A4.23.2

Distribution of answers to the item M032116

Condition	Answers				Overall
	A	B (correct)	C	D	
Regular (attentive)	82 (16.5%)	299 (60.3%)	93 (18.8%)	22 (4.4%)	496 (100.0%)
Decreased motivation	122 (25.1%)	242 (49.8%)	96 (19.8%)	26 (5.3%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 14.13$, $p = 0.003$

Answering time

Figure A4.23.3

Distribution of response time to the item M032116

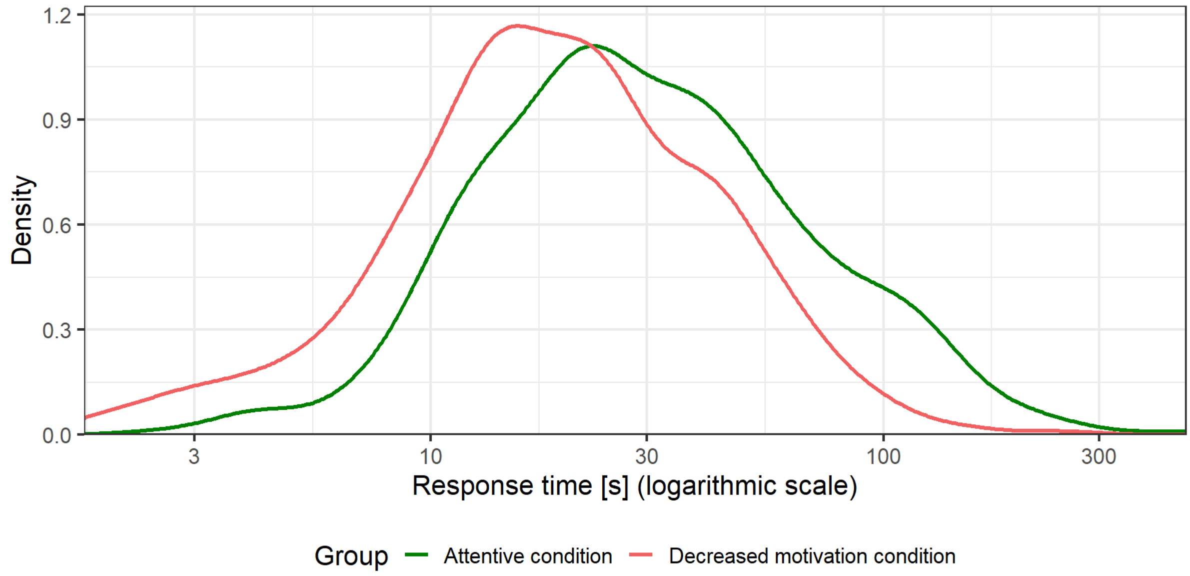


Table A4.23.3

Median and mean log response time to the item M032116

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	42.0	3.39
Decreased motivation	25.3	2.92
ANOVA for mean log(response time) differences: *F*(1, 980) = 82.06, *p* = 0.000		

gmm2016z01 (item set 2)

ICCs and fit statistics

Figure A4.24.1

Item gmm2016z01 item characteristics curves (ICCs)

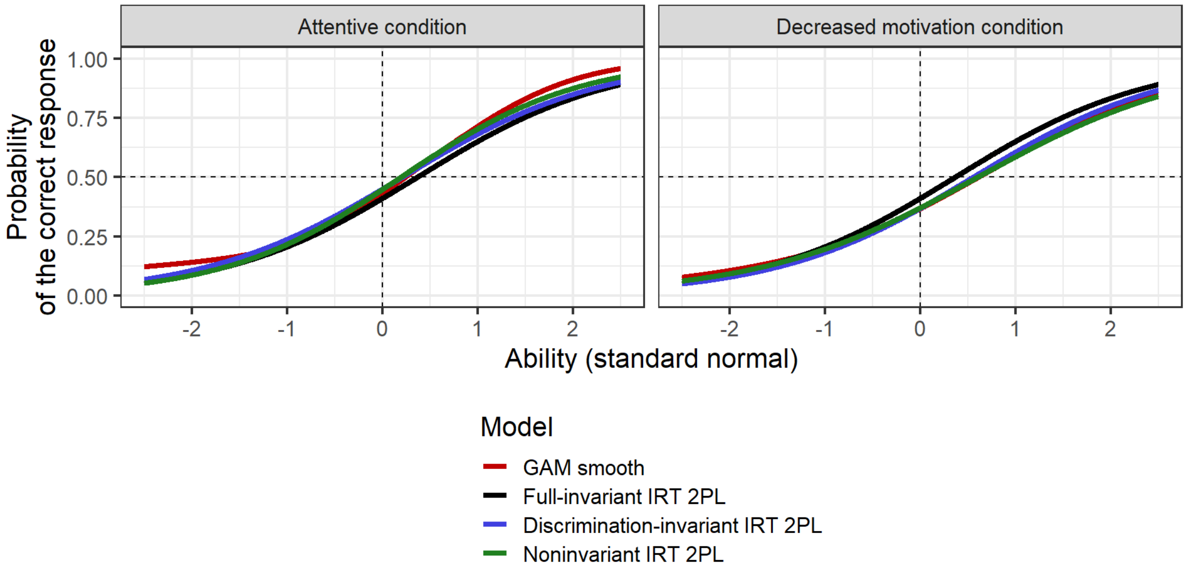


Table A4.24.1

Item gmm2016z01 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.99	0.37−		
Attentive condition			0.05	0.009
Decreased motivation condition			0.00	0.662
Discrimination-invariant 2PL model				
All groups	0.97			
Attentive condition		0.21	0.05	0.007
Decreased motivation condition		0.56	0.00	0.671
Noninvariant 2PL model				
Attentive condition	1.08	0.20	0.05	0.010
Decreased motivation condition	0.88	0.60	0.00	0.689

Frequency of answers

Figure A4.24.2

Distribution of answers to the item gmm2016z01

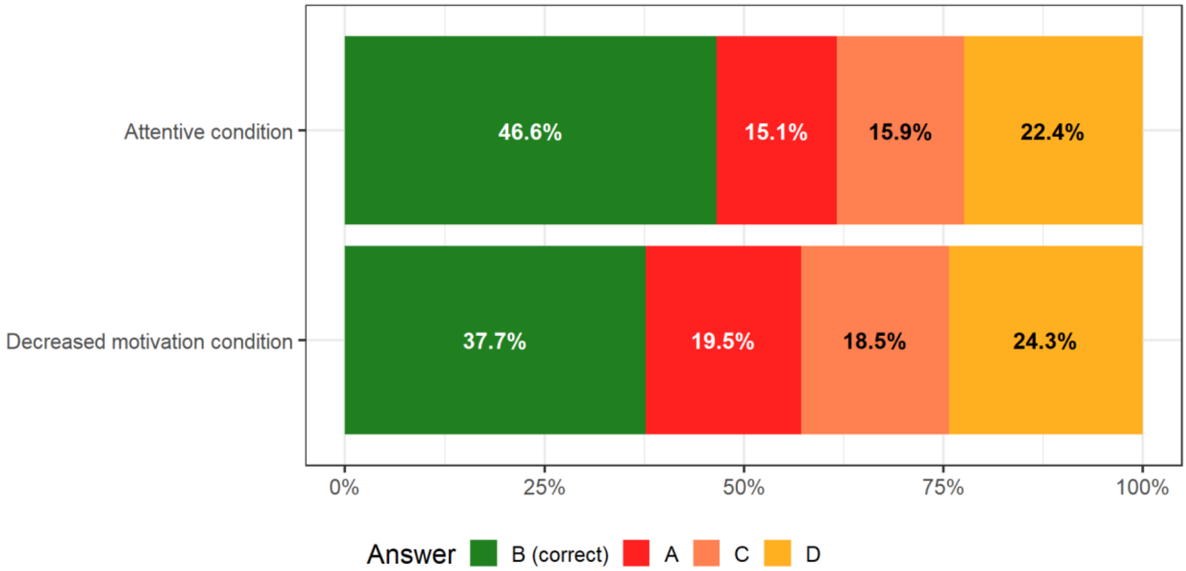


Table A4.24.2

Distribution of answers to the item gmm2016z01

Condition	Answers				Overall
	A	B (correct)	C	D	
Regular (attentive)	75 (15.1%)	231 (46.6%)	79 (15.9%)	111 (22.4%)	496 (100.0%)
Decreased motivation	95 (19.5%)	183 (37.7%)	90 (18.5%)	118 (24.3%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 8.75$, $p = 0.033$

Answering time

Figure A4.24.3

Distribution of response time to the item gmm2016z01

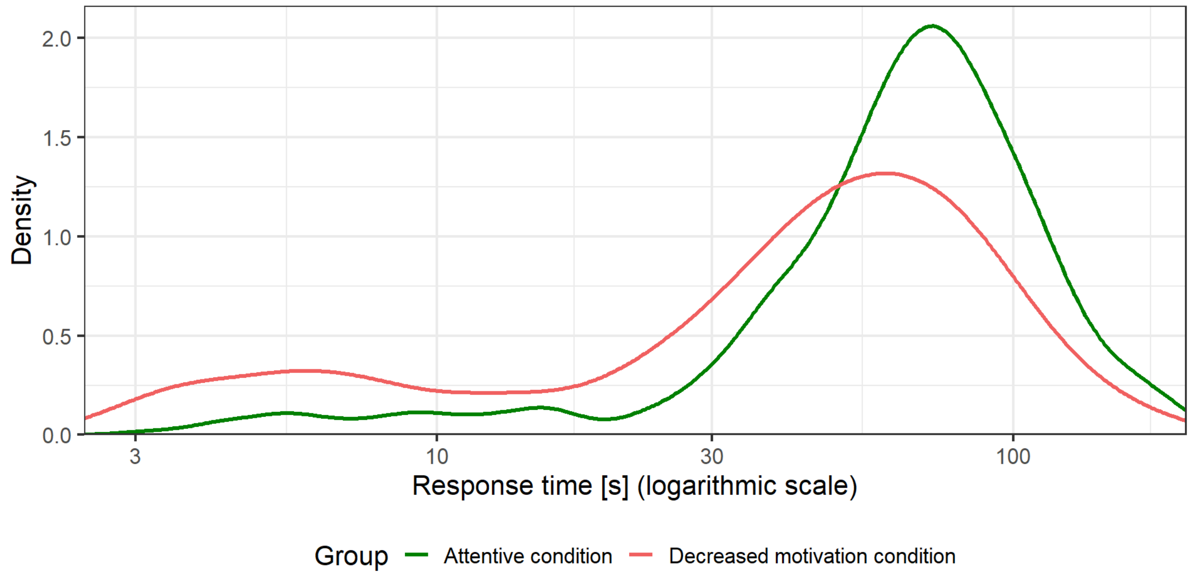


Table A4.24.3

Median and mean log response time to the item gmm2016z01

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	71.0	4.09
Decreased motivation	51.0	3.56
ANOVA for mean log(response time) differences: *F*(1, 980) = 88.80, *p* = 0.000		

M042236 (item set 2)

ICCs and fit statistics

Figure A4.25.1

Item M042236 item characteristics curves (ICCs)

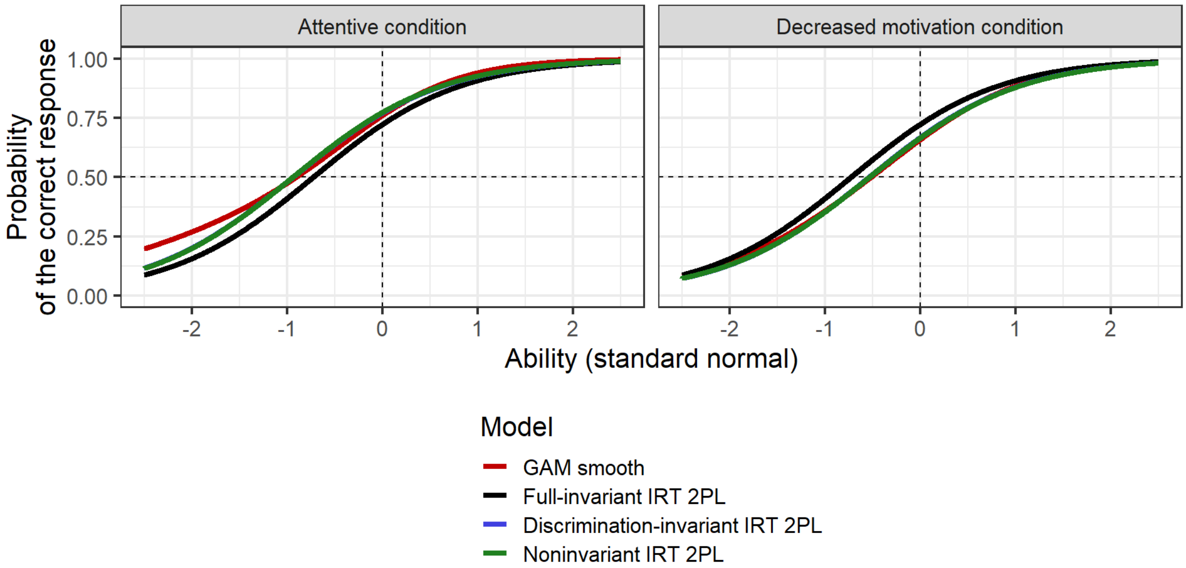


Table A4.25.1

Item M042236 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.32	−0.72		
Attentive condition			0.05	0.012
Decreased motivation condition			0.05	0.007
Discrimination-invariant 2PL model				
All groups	1.30			
Attentive condition		−0.94	0.03	0.174
Decreased motivation condition		−0.53	0.03	0.165
Noninvariant 2PL model				
Attentive condition	1.31.	−0.94	0.04	0.044
Decreased motivation condition	1.29	−0.53	0.03	0.114

Frequency of answers

Figure A4.25.2

Distribution of answers to the item M042236

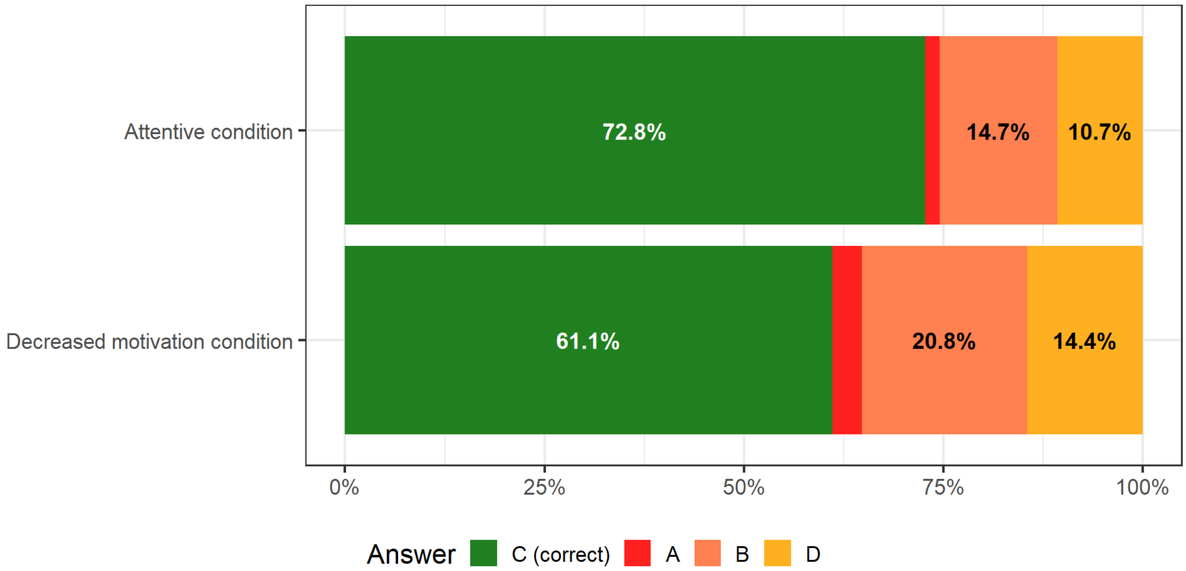


Table A4.25.2

Distribution of answers to the item M042236

Condition	Answers				Overall
	A	B	C (correct)	D	
Regular (attentive)	9 (1.8%)	73 (14.7%)	361 (72.8%)	53 (10.7%)	496 (100.0%)
Decreased motivation	18 (3.7%)	101 (20.8%)	297 (61.1%)	70 (14.4%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 15.98$, $p = 0.001$

Answering time

Figure A4.25.3

Distribution of response time to the item M042236

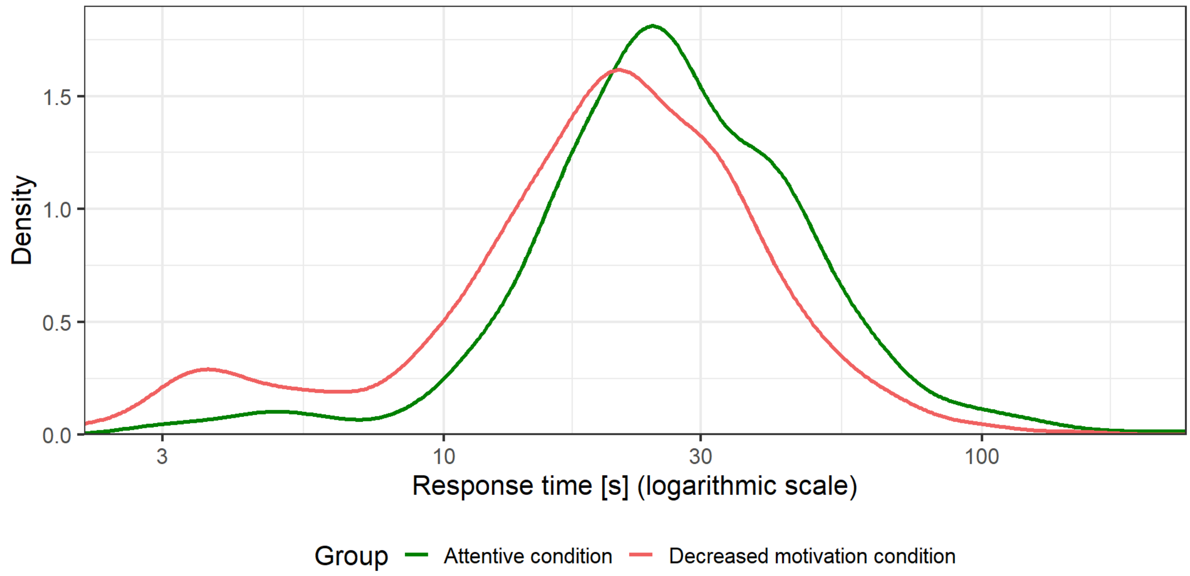


Table A4.25.3

Median and mean log response time to the item M042236

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	31.2	3.26
Decreased motivation	23.6	2.94
ANOVA for mean log(response time) differences: *F*(1, 980) = 56.02, *p* = 0.000		

gm2002z05 (item set 2)

ICCs and fit statistics

Figure A4.26.1

Item gm2002z05 item characteristics curves (ICCs)

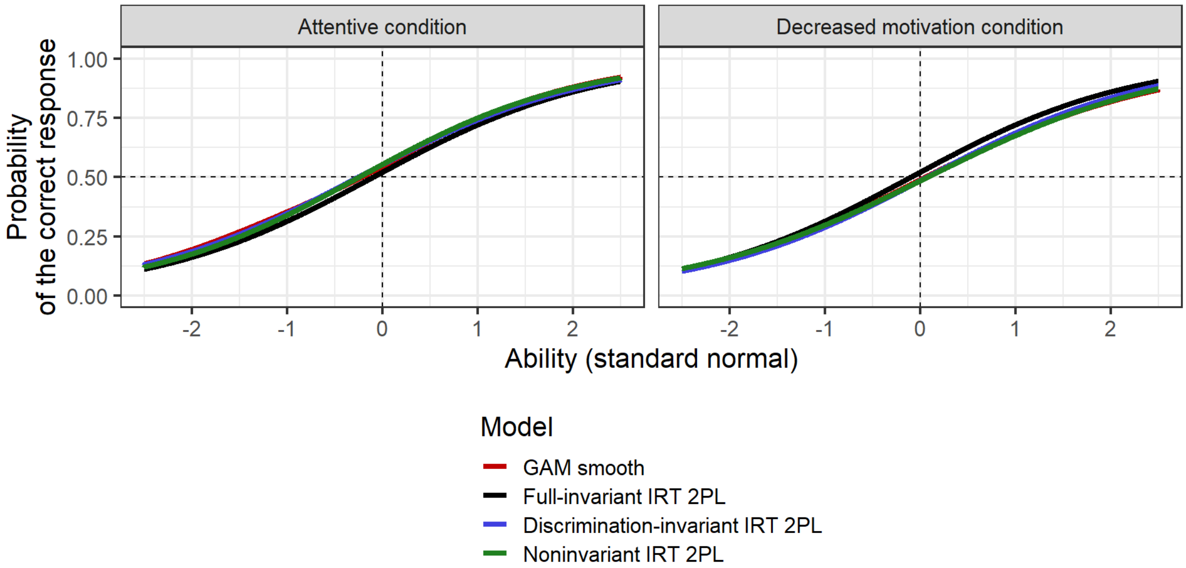


Table A4.26.1

Item gm2002z05 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.86	−0.10		
Attentive condition			0.05	0.007
Decreased motivation condition			0.00	0.608
Discrimination-invariant 2PL model				
All groups	0.84			
Attentive condition		−0.25	0.04	0.091
Decreased motivation condition		0.07	0.00	0.785
Noninvariant 2PL model				
Attentive condition	0.88	−0.25	0.03	0.178
Decreased motivation condition	0.80	0.08	0.00	0.679

Frequency of answers

Figure A4.26.2

Distribution of answers to the item gm2002z05

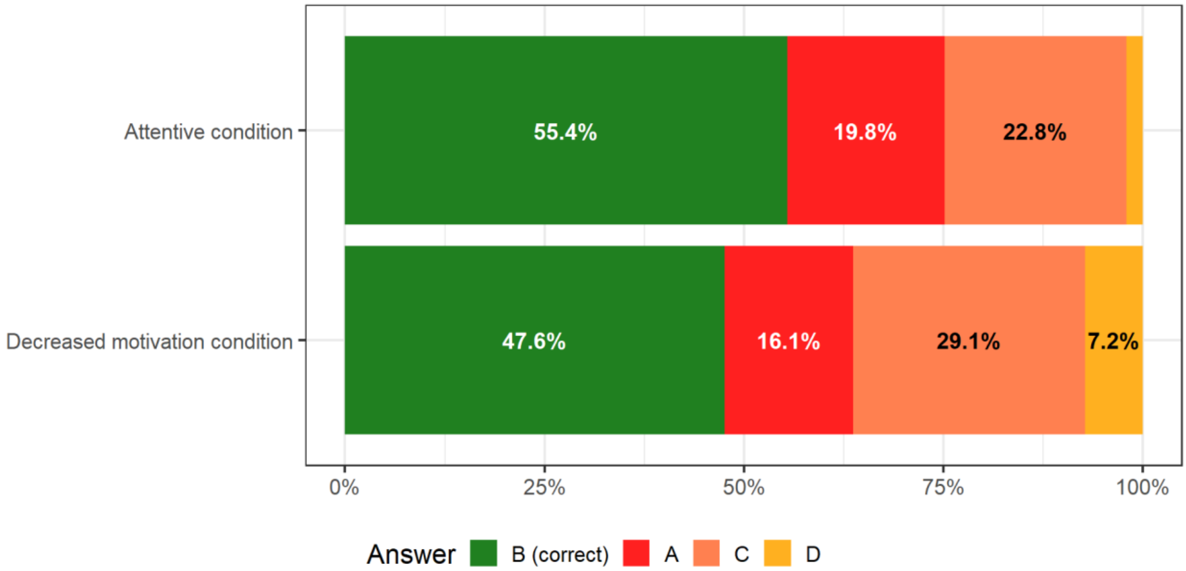


Table A4.26.2

Distribution of answers to the item gm2002z05

Condition	Answers				Overall
	A	B (correct)	C	D	
Regular (attentive)	98 (19.8%)	275 (55.4%)	113 (22.8%)	10 (2.0%)	496 (100.0%)
Decreased motivation	78 (16.1%)	231 (47.6%)	141 (29.1%)	35 (7.2%)	485 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 22.95$, $p = 0.000$

Answering time

Figure A4.26.3

Distribution of response time to the item gm2002z05

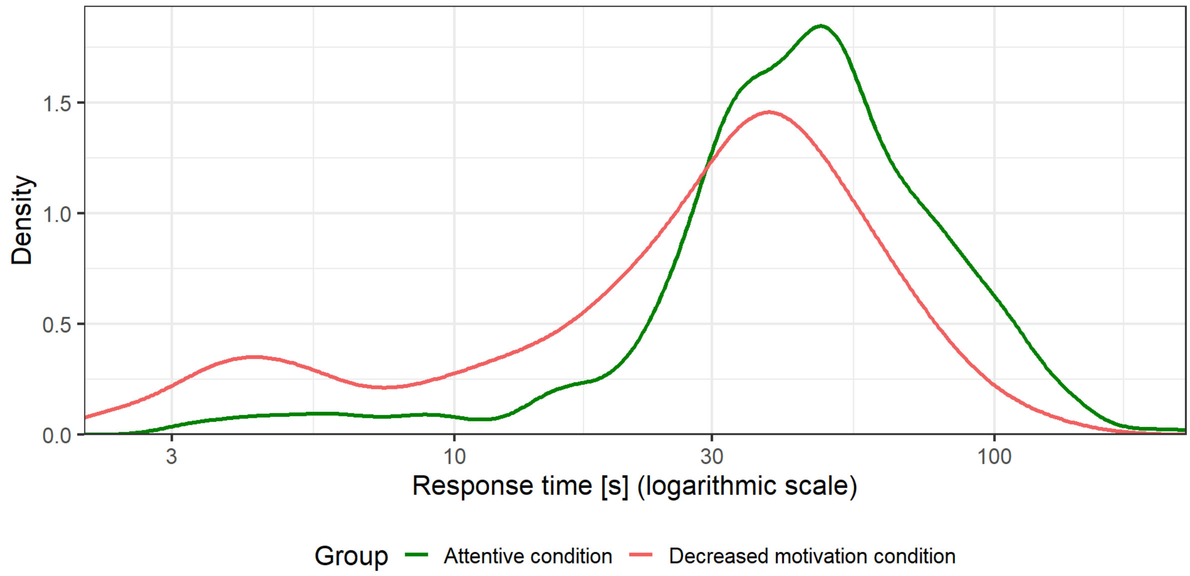


Table A4.26.3

Median and mean log response time to the item gm2002z05

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	50.6	3.75
Decreased motivation	34.2	3.22
ANOVA for mean log(response time) differences: $F(1, 980) = 109.67$, $p = 0.000$		

gmm2016z03 (item set 2)

ICCs and fit statistics

Figure A4.27.1

Item gmm2016z03 item characteristics curves (ICCs)

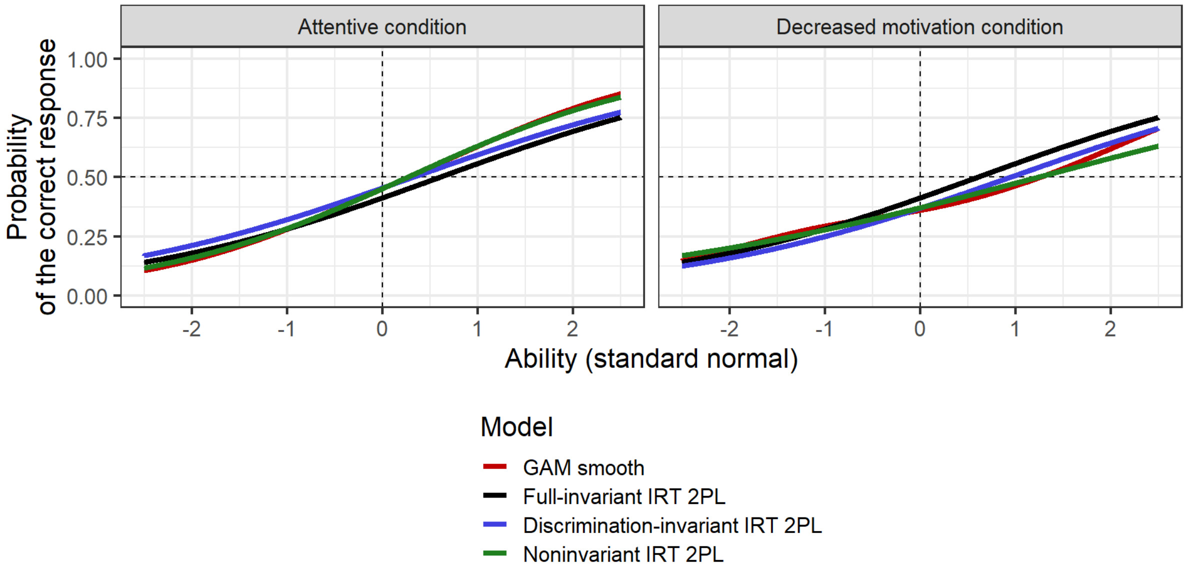


Table A4.27.1

Item gmm2016z03 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	0.58	0.60		
Attentive condition			0.04	0.048
Decreased motivation condition			0.03	0.190
Discrimination-invariant 2PL model				
All groups	0.57			
Attentive condition		0.32	0.01	0.391
Decreased motivation condition		0.95	0.03	0.141
Noninvariant 2PL model				
Attentive condition	0.73	0.27	0.02	0.224
Decreased motivation condition	0.43	1.24	0.03	0.164

Frequency of answers

Figure A4.27.2

Distribution of answers to the item gmm2016z03

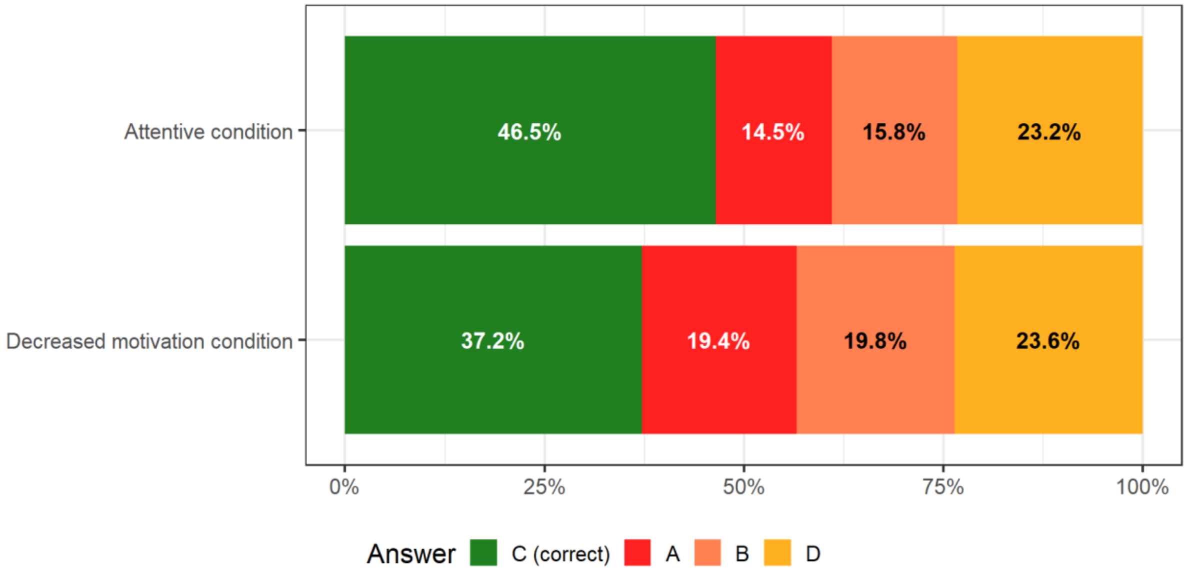


Table A4.27.2

Distribution of answers to the item gmm2016z03

Condition	Answers				Overall
	A	B	C (correct)	D	
Regular (attentive)	72 (14.5%)	78 (15.8%)	230 (46.5%)	115 (23.2%)	495 (100.0%)
Decreased motivation	94 (19.4%)	96 (19.8%)	180 (37.2%)	114 (23.6%)	484 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 10.76$, $p = 0.013$

Answering time

Figure A4.27.3

Distribution of response time to the item gmm2016z03

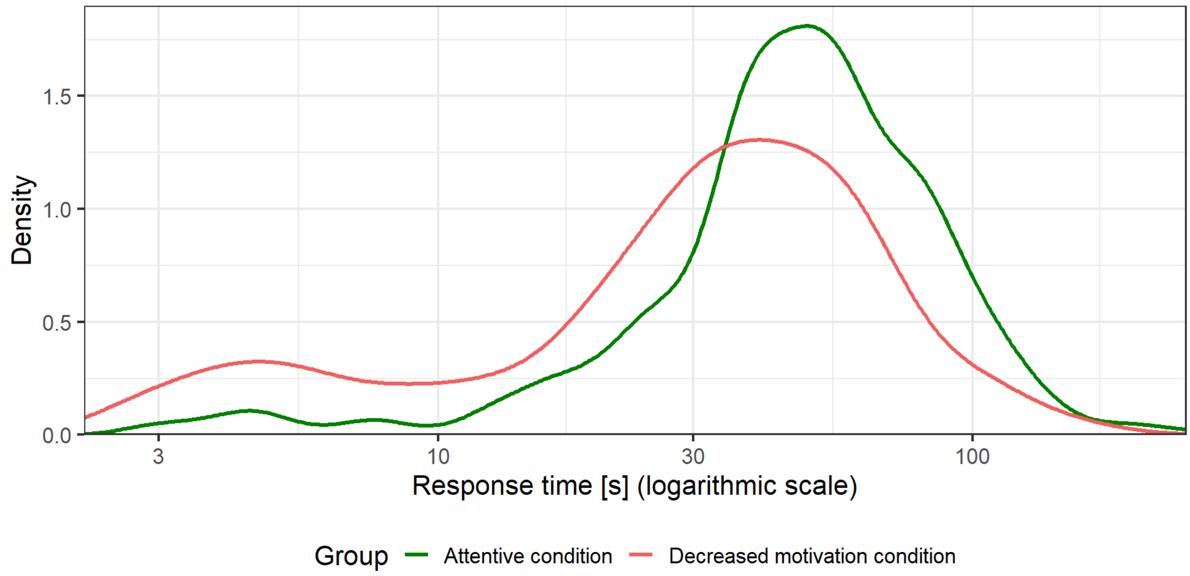


Table A4.27.3

Median and mean log response time to the item gmm2016z03

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	53.9	3.80
Decreased motivation	38.1	3.31
ANOVA for mean log(response time) differences: *F*(1, 980) = 91.35, *p* = 0.000		

M022101 (item set 2)

ICCs and fit statistics

Figure A4.28.1

Item M022101 item characteristics curves (ICCs)

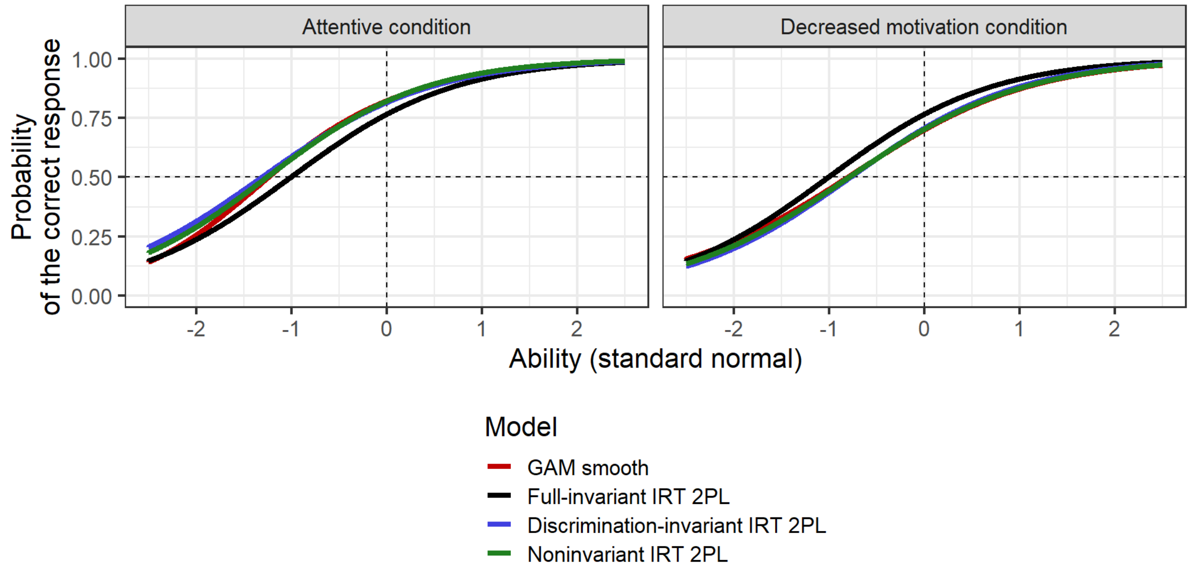


Table A4.28.1

Item M022101 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.18	−1.01		
Attentive condition			0.00	0.449
Decreased motivation condition			0.03	0.114
Discrimination-invariant 2PL model				
All groups	1.14			
Attentive condition		−1.31	0.03	0.093
Decreased motivation condition		−0.77	0.00	0.733
Noninvariant 2PL model				
Attentive condition	1.21	−1.26	0.00	0.540
Decreased motivation condition	1.09	−0.79	0.00	0.598

Frequency of answers

Figure A4.28.2

Distribution of answers to the item M022101

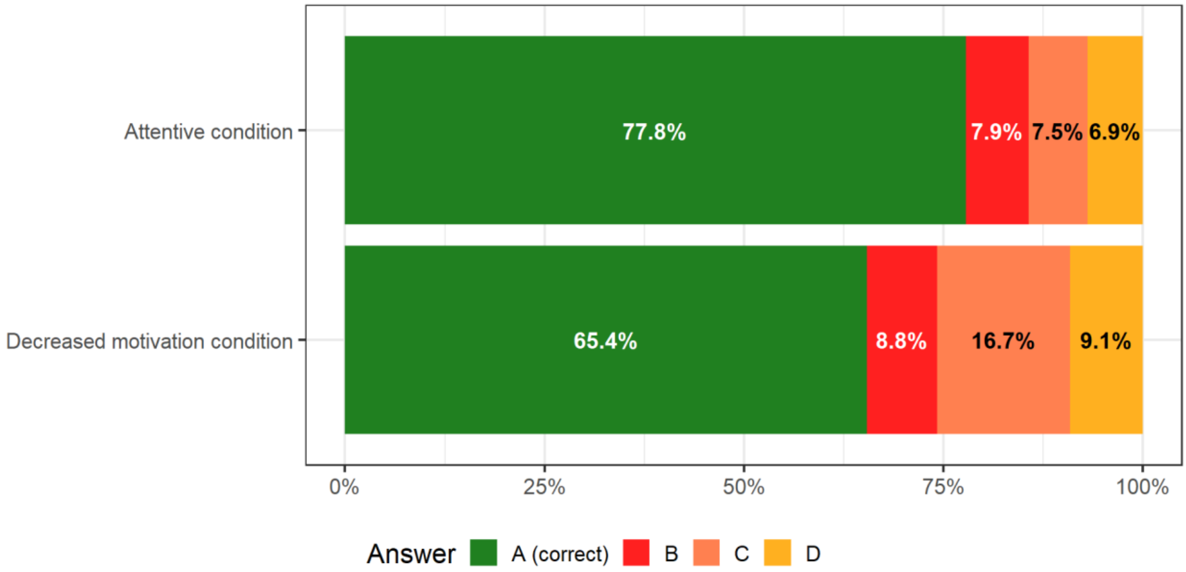


Table A4.28.2

Distribution of answers to the item M022101

Condition	Answers				Overall
	A (correct)	**B**	**C**	**D**	
Regular (attentive)	386 (77.8%)	39 (7.9%)	37 (7.5%)	34 (6.9%)	496 (100.0%)
Decreased motivation	318 (65.4%)	43 (8.8%)	81 (16.7%)	44 (9.1%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 24.35$, $p = 0.000$

Answering time

Figure A4.28.3

Distribution of response time to the item M022101

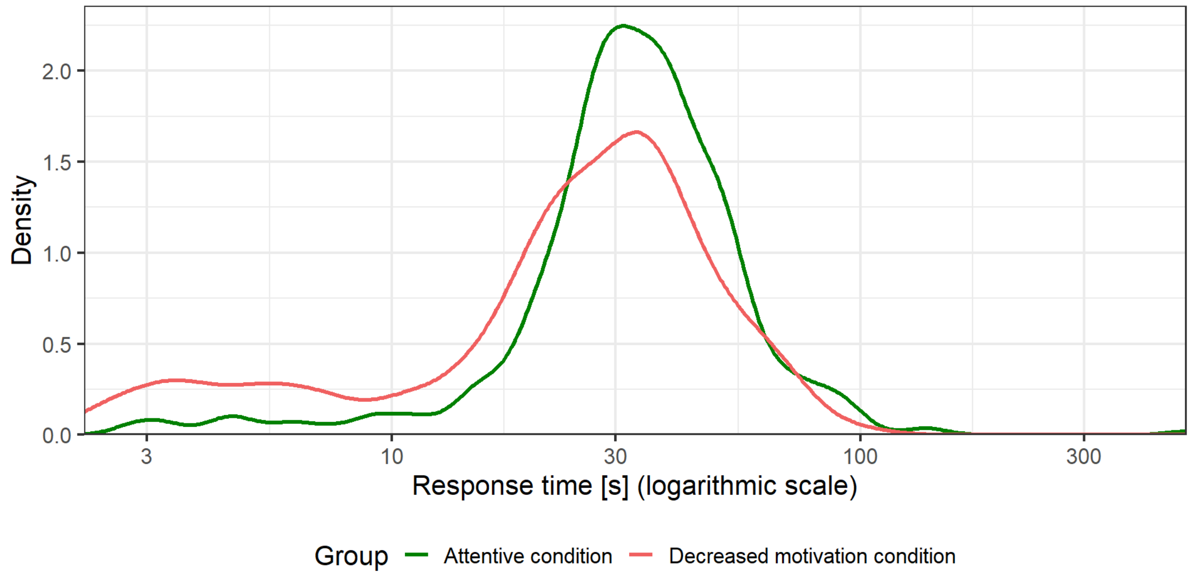


Table A4.28.3

Median and mean log response time to the item M022101

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	36.9	3.45
Decreased motivation	28.7	3.09
ANOVA for mean log(response time) differences: *F*(1, 980) = 61.39, *p* = 0.000		

M052084 (item set 2)

ICCs and fit statistics

Figure A4.29.1

Item M052084 item characteristics curves (ICCs)

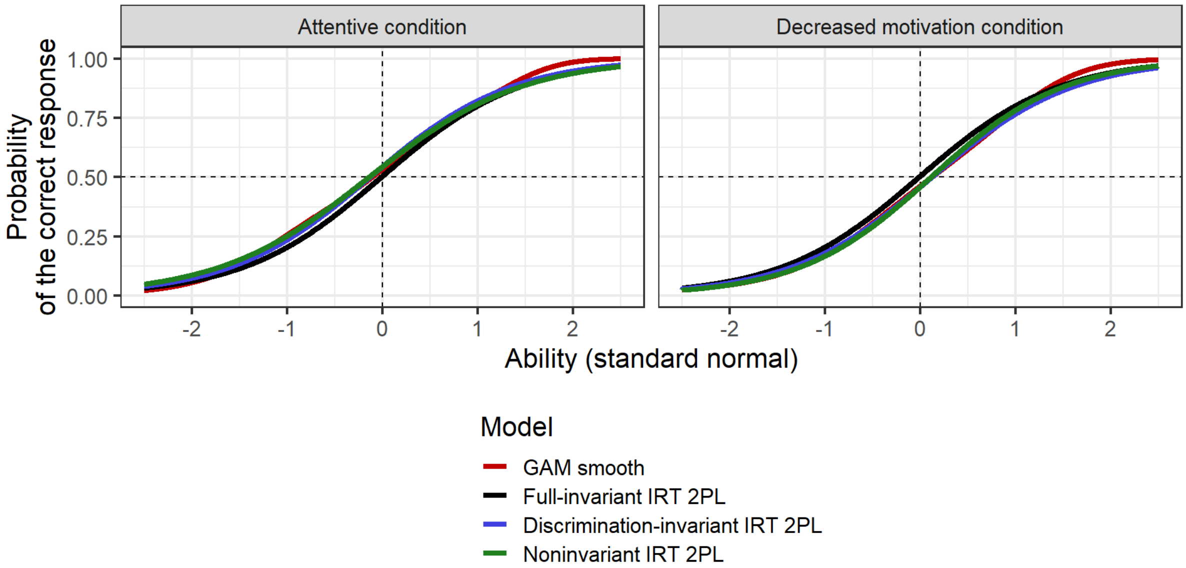


Table A4.29.1

Item M052084 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.38	−0.01		
Attentive condition			0.03	0.176
Decreased motivation condition			0.05	0.003
Discrimination-invariant 2PL model				
All groups	1.35			
Attentive condition		−0.13	0.04	0.087
Decreased motivation condition		0.12	0.05	0.015
Noninvariant 2PL model				
Attentive condition	1.27	−0.14	0.04	0.041
Decreased motivation condition	1.44	0.11	0.03	0.093

Frequency of answers

Figure A4.29.2

Distribution of answers to the item M052084

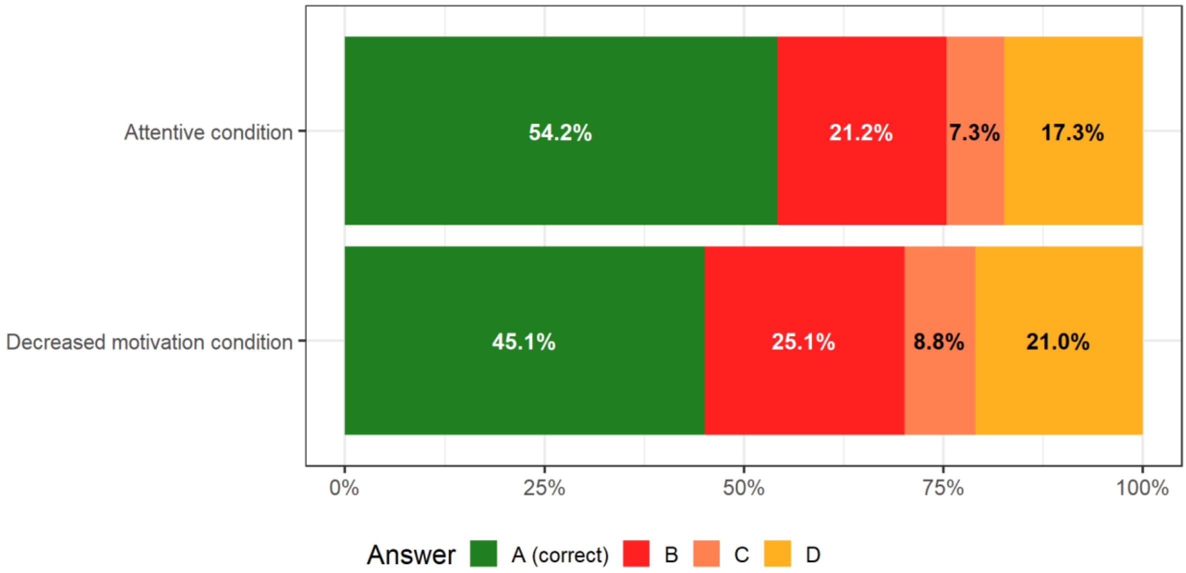


Table A4.29.2

Distribution of answers to the item M052084

Condition	Answers				Overall
	A (correct)	B	C	D	
Regular (attentive)	269 (54.2%)	105 (21.2%)	36 (7.3%)	86 (17.3%)	496 (100.0%)
Decreased motivation	219 (45.1%)	122 (25.1%)	43 (8.8%)	102 (21.0%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 8.28$, $p = 0.041$

Answering time

Figure A4.29.3

Distribution of response time to the item M052084

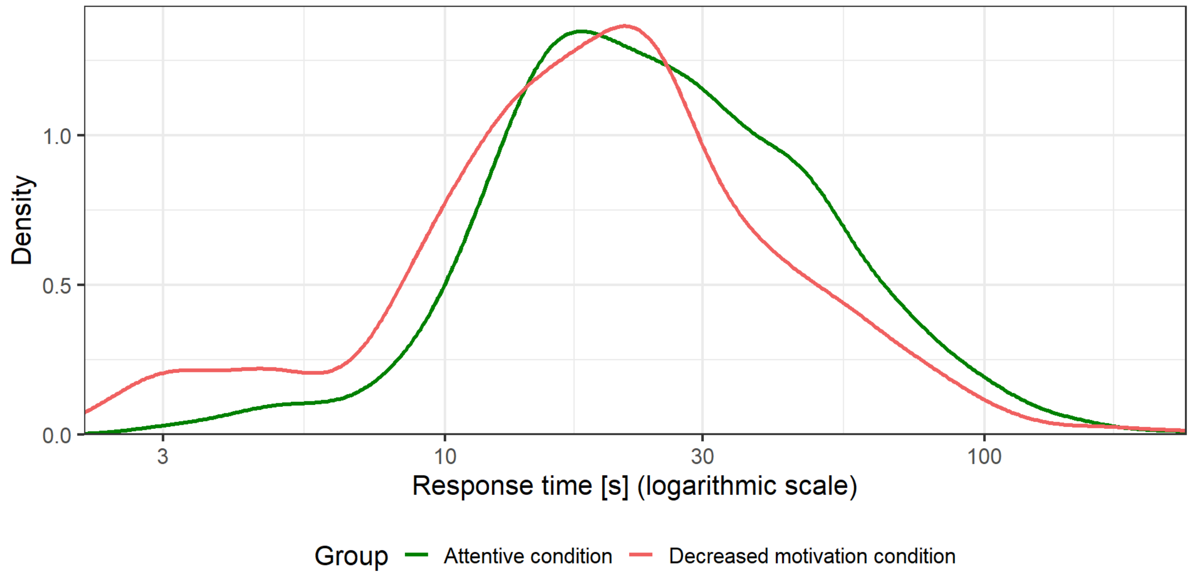


Table A4.29.3

Median and mean log response time to the item M052084

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	31.2	3.21
Decreased motivation	25.2	2.92
ANOVA for mean log(response time) differences: $F(1, 980) = 36.76$, $p = 0.000$		

M032094 (item set 2)

ICCs and fit statistics

Figure A4.30.1

Item M032094 item characteristics curves (ICCs)

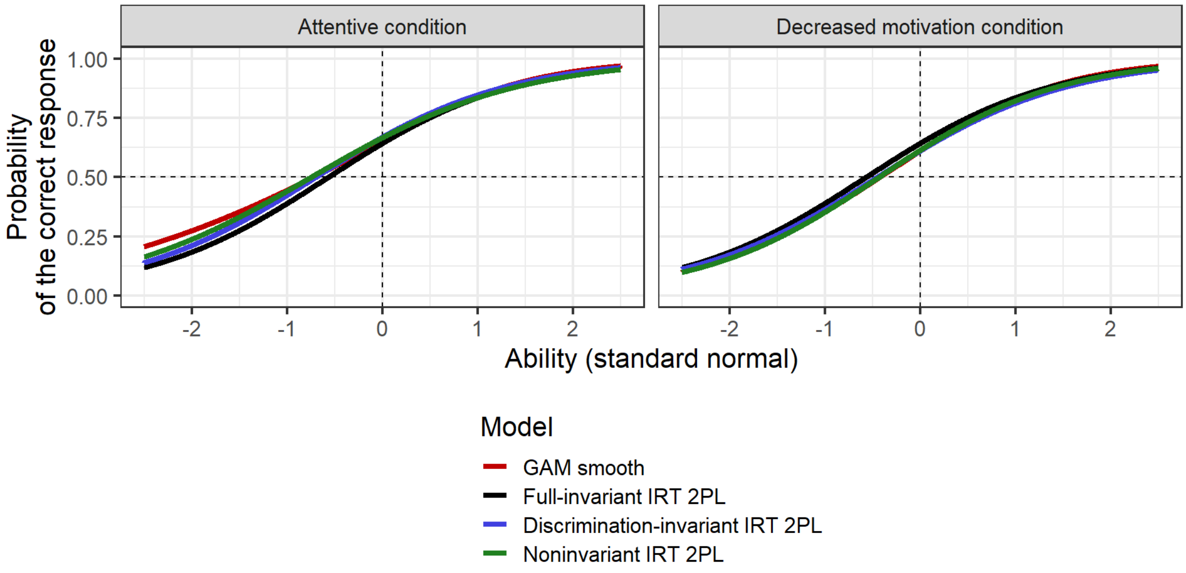


Table A4.30.1

Item M032094 parameters and RMSEA in different model specifications

	Discrimination (a)	Difficulty (b)	RMSEA	p
Full-invariant 2PL model				
All groups	1.04	−0.56		
Attentive condition			0.05	0.023
Decreased motivation condition			0.02	0.205
Discrimination-invariant 2PL model				
All groups	1.01			
Attentive condition		−0.70	0.02	0.224
Decreased motivation condition		−0.45	0.03	0.119
Noninvariant 2PL model				
Attentive condition	0.93	−0.74	0.03	0.161
Decreased motivation condition	1.07	−0.43	0.02	0.216

Frequency of answers

Figure A4.30.2

Distribution of answers to the item M032094

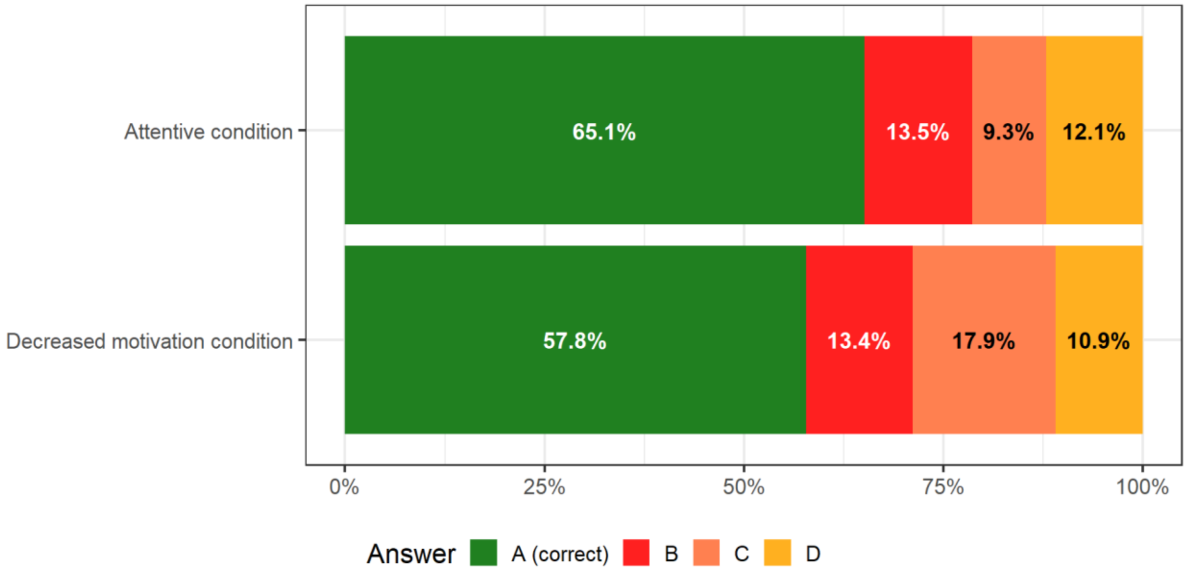


Table A4.30.2

Distribution of answers to the item M032094

Condition	Answers				Overall
	A (correct)	**B**	**C**	**D**	
Regular (attentive)	323 (65.1%)	67 (13.5%)	46 (9.3%)	60 (12.1%)	496 (100.0%)
Decreased motivation	281 (57.8%)	65 (13.4%)	87 (17.9%)	53 (10.9%)	486 (100.0%)

Pearson Chi-squared tests of independence: $\chi^2(3) = 15.92$, $p = 0.001$

Answering time

Figure A4.30.3

Distribution of response time to the item M032094

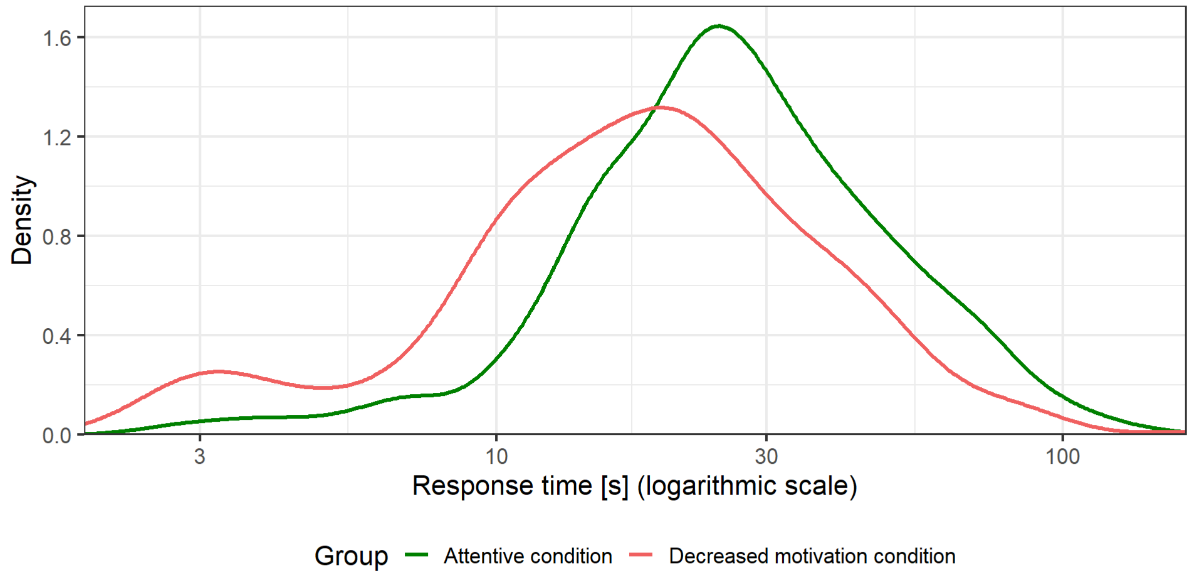


Table A4.30.3

Median and mean log response time to the item M032094

Condition	Median response time [s]	Mean log(response time)
Regular (attentive)	31.1	3.24
Decreased motivation	22.4	2.84
ANOVA for mean log(response time) differences: $F(1, 980) = 78.57$, $p = 0.000$		

Annex 5

A5.1 File naming convention

File names were constructed using the convention:

`seqN_(eq|uneq)_(long|wide)_ID`

where:

- `N` is the number of AAoIs in the AAoIs grid;
- `(eq|uneq)` – "eq" for files with equally spaced time grid or "uneq" for files recording all crossings between the AAoIs;
- `(long|wide)` – "long" for files describing a single respondent-item or or "wide" for files describing a single respondent;
- `ID` is a reshuffled respondent's (*wide* format) or respondent-screen's (*long* format) ID

A5.1.1 Datasets storing individual respondent-items

- Each file represents a single respondent-item
- Time points are represented by columns
- Information about status and auxiliary information are represented by rows

Data with equally spaced time grid

- Different time intervals (of the same length) are represented by columns

For a grid of `n` AAoIs:

- The first `n` rows store binary information about whether the cursor was within a given AAoI at the end of a given second (0 – no, 1 – yes)
  - Within each column values of these rows sum up to 1
- Row `n+1` stores binary information about whether an item was stemless (0 – no, 1 – yes)
- Row `n+2` stores binary information about whether an item was in the testlet which was presented first to a given test-taker (0 – yes, 1 – no)

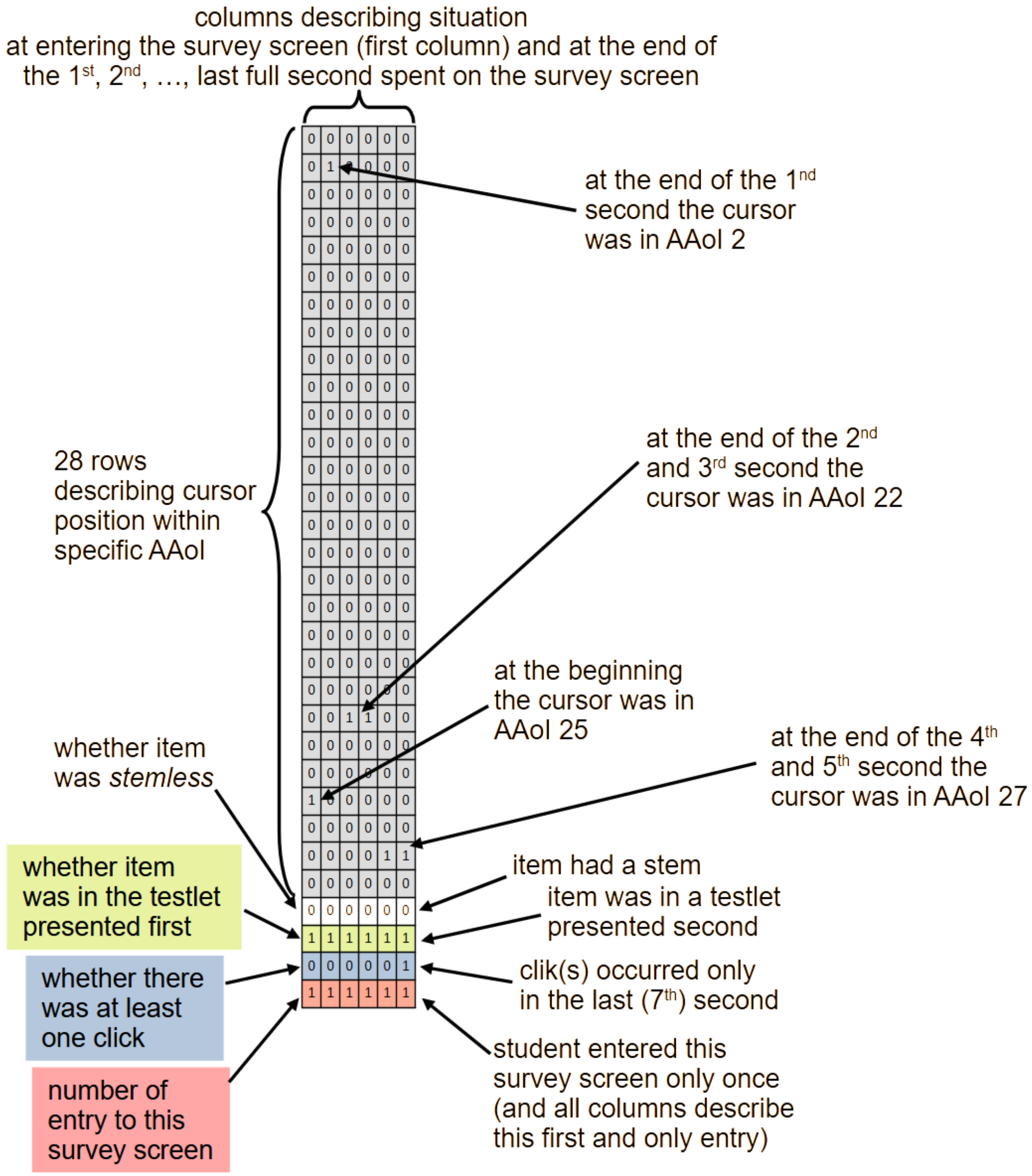
- Row `n+3` stores binary information about whether there was (at least one) click during a given second (0 – no, 1 – yes)

- Row `n+4` stores an integer describing which entry to the survey screen containing the item it was (1 – first entry, 2 – first re-entry, 3 – second re-entry, ...)

Structure of such a file was presented in Figure A5.1.1 below. In this example a grid of 28 AAoIs was used (see Figure 4 to determine the area covered by the specific AAoIs) and respondent spent less than six seconds on this survey screen.

Figure A5.1.1

Structure of the dataset describing respondent's actions on a single survey screen using the equally spaced time grid



- Different respondent actions (crossing to another AAoI or click) are represented by columns

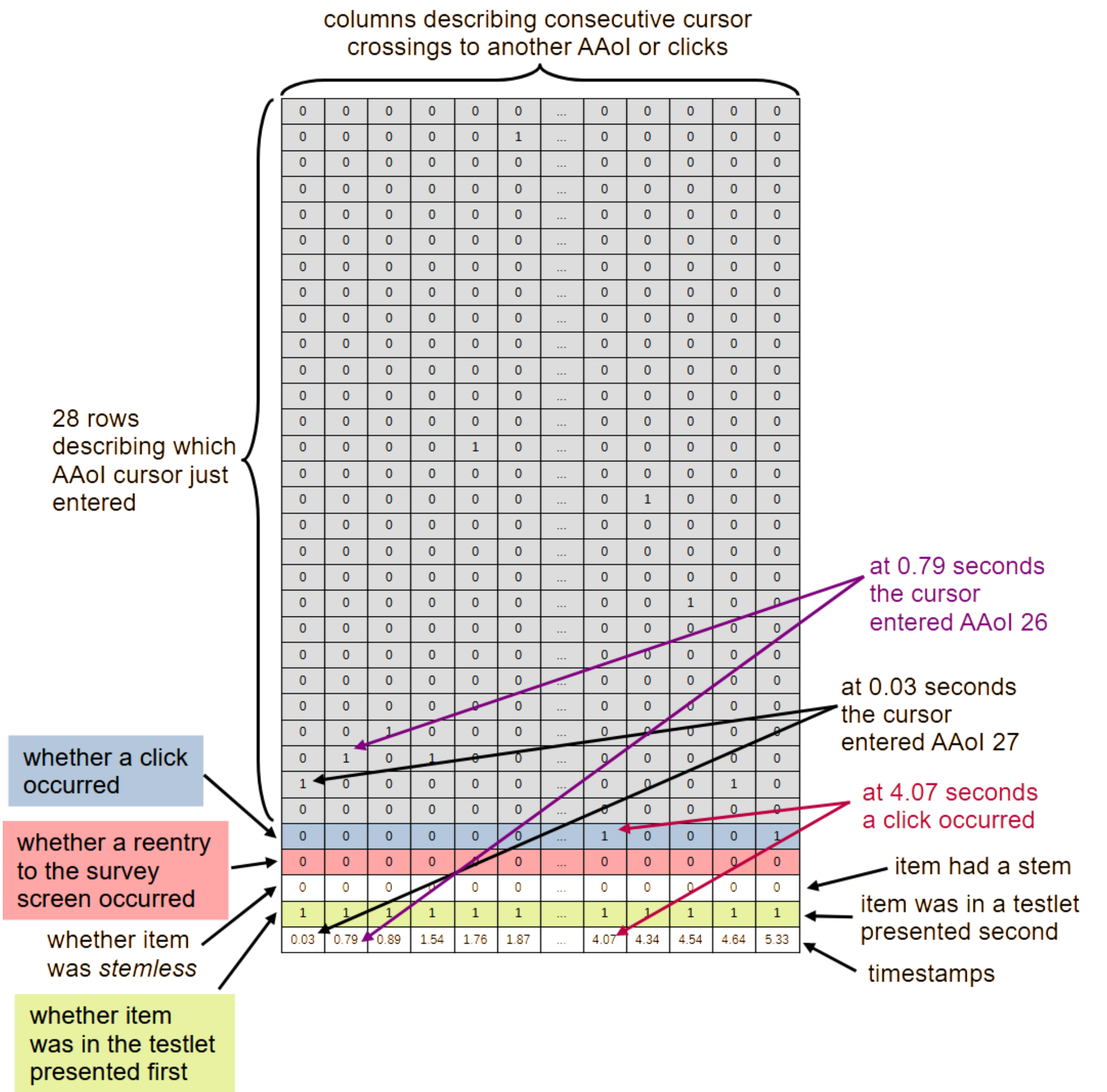
For a grid of `n` AAoIs:

- The first `n` rows store binary information about whether cursor just entered a given AAoI (0 - no, 1 - yes)
- Row `n+1` stores binary information whether a click occurred (0 – no, i.e. column describes cursor passing from one AAoI to another, 1 – yes)
    - In the case of clicks, there is no information within which AAoI it occurred
- Row `n+2` stores binary information whether a re-entry to the item occurred (0 – no, 1 – yes)
    - Within each column values of the first `n+2` rows sum up to 1
- Row `n+3` stores binary information about whether an item was stemless (0 – no, 1 – yes)
- Row `n+4` binary information about whether an item was in the testlet which was presented first to a given test-taker (0 – no, 1 – yes)
- Row `n+5` stores the time since the survey screen loaded to a given event (in seconds)

Structure of such a file was presented in Figure A5.1.2 below. In this example a grid of 28 AAoIs was used (see Figure 4 to determine the area covered by the specific AAoIs). The file describes the file for the same respondent and screen as in Figure A5.1.1.

Figure A5.1.2

Structure of the dataset describing respondent's actions on a single survey screen using the unequally spaced time grid



Data with unequally spaced time grid was used for all predictive models. Initial analysis showed that this structure outperform equally spaced time. However we produced both data structures for further elaborations.

A5.2 Datasets storing individual respondents

Data format is basically the same as in datasets storing individual respondent-items with the following exceptions:

- Each file represents a single respondent – across all the items for which log-data collection succeeded for a given respondent.
- Time points from consecutive items are stored in the subsequent columns.
- In files storing information about all crossings between AAoIs and clicks, a row storing timestamp of a given event is constructed in a way that for each consecutive item the sum of screen times from the previous items (for which log-data collection succeeded for a given respondent) is added to it, so that the values in this row increase from left to right.
    - Please note, that since this procedure only considers those items for which log-data collection succeeded for a given respondent, this means that the data indirectly includes information about whether any technical difficulties occurred (resulting in the failure of log-data collection) during the survey for a given respondent.
- There are additional rows added at the bottom of each dataset describing which item a given column refers to. These are binary coded (0 - not a given screen, 1 – a given screen):
    - There are 31 such rows describing items in the following order: 1) M032579, 2) M042254, 3) M042222, 4) gm2006z08, 5) M042260, 6) gmm2015z07, 7) M032295, 8) M042031, 9) gm2004z04, 10) M032679, 11) M032198, 12) gmm2014z11, 13) M042041, 14) e82022z7, 15) M032166, 16) M042032, 17) M052302, 18) gmm2016z20, 19) M032738, 20) M032477, 21) gmm2018z07, 22) M032704, 23) M032116, 24) gmm2016z01, 25) M042236, 26) gm2002z05, 27) gmm2016z03, 28) M022101, 29) M052084, 30) M032094, 31) *bad* item
    - Within each column values of these 31 rows sum up to 1.

A5.3 Datasets with auxiliary information

We prepared datasets with auxiliary information to enable various statistical analyses, including the examination of the effectiveness of experimental stimuli and psychometric analyses (see Tables A5.1 and A5.2).

Table A5.1

Auxiliary Information for Files Describing Separate Respondent-items

Variable	Format	Description
idRespScreen	integer: >0	respondent-screen's reshuffled ID
token	string	respondent's ID
study	integer: {0, 1}	in which part of the study the test was taken: 0 – main, 1 – pilot
status	integer: {0, 1}	quality control (QC) status of the test: 0 – passed QC, 1 – failed QC
itemSetsOrder	integer: {0, 1}	order of presentation of testlets during the test: 0 – testlet one first, 1 – testlet two first
instructionOrder	integer: {0, 1}	order of instruction during the test: 0 – attentive instruction first, 1 – inattentive instruction first
localitySize	integer: {0, 1, 2}	school location size: 0 – village, 1 – small town, 2 – big town/city
grade	integer: {0, 1}	test-taker's grade: 0 – seventh, 1 – eighth
gender	integer: {0, 1, 2}	test-taker's gender: 0 – male, 1 – female
motivation	integer: {0, 1, 2, 3}	test-taker's self declared motivation ("How hard did you try during the test?"): 0 – Not at all, 1 – Not much, 2 –Rather hard, 3 – Very hard
strategy2	integer: {0, 1, 2, 3, 4}	test-taker's self declared response strategy under the inattentive instruction: 0 – solving only easy items, 1 – informed guessing, 2 – attentive answering (indicates a failure of manipulation), 3 – speeding, 4 – other
books	integer: {0, 1, 2, 3, 4}	number of books in the test-taker's home (self-declared): 0 – 0-10, 1 – 11-25, 2 – 26-100, 3 – 101-200, 4 – >200
mouseUse	integer: {0, 1, 2, 3}	test taker's self declared way of using a mouse: 0 – "tracking what I read", 1 – "marking answers only", 2 – other "unrelated moves" or 3- missing answer
scoreAtt	numeric	standardized test-taker's sum score of the testlet with attentive instruction
scoreInatt	numeric	standardized test-taker's sum score of the testlet with inattentive instruction
scoreAttTercile	integer: {0, 1, 2}	tercile of test-taker's sum score of the testlet with attentive instruction
scoreInattTercile	integer: {0, 1, 2}	tercile of test-taker's sum score of the testlet with inattentive instruction
screen	string	item (survey screen name)
attentive	integer: {0, 1}	according to what instruction did given test-taker answered a given item: 0 – inattentive, 1 – attentive

Variable	Format	Description
order	integer: {0, 1}	whether a given item was in the testlet presented first or in the testlet presented second to a given test-taker: 0 – in the testlet presented first, 1 – in the testlet presented second
noHead	integer: {0, 1}	whether a given item was "headless": 0 – no (i.e. item has a "head"), 1 – yes (i.e. item has no "head")
itemSet	integer: {1, 2}	which testlet does a given item belong to: 1 – testlet one, 2 – testlet two, 3 – "bad" item
badItem	integer: {0, 1}	which version of the "bad" item was answered by a given test-taker: 0 – version with no defect, 1 – version with a fault (i.e. many correct answers)
answer	integer: {0, 1, 2, 3, 4} | missing	answer of a given test taker on a given item: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D, 4 – answer E (the last one applies only to some items)
score	integer: {0, 1}	score of a given test taker on a given item: 0 – incorrect answer or missing answer, 1 – correct answer
screenTime	numeric: >0	survey screen (item) time in seconds
screenTimeLog	numeric	natural logarithm of survey screen time
log_dX	numeric	natural logarithm of the total distance (in pixels) travelled by the cursor in the horizontal direction on a given survey screen
sqrt_flipsX	numeric: ≥0	square root of the number of flips (changes in the cursor move direction) in the horizontal axis on a given survey screen
log_dY	numeric	natural logarithm of the total distance (in pixels) travelled by the cursor in the vertical direction on a given survey screen
sqrt_flipsY	numeric: ≥0	square root of the number of flips (changed of the cursor move direction) in the vertical axis on a given survey screen
log_vX	numeric	natural logarithm of the average cursor speed (in pixels per second) in the horizontal axis on a given survey screen
log_vY	numeric	natural logarithm of the average cursor speed (in pixels per second) in the vertical axis on a given survey screen
log_aX	numeric	natural logarithm of the average absolute cursor acceleration (in pixels per second squared) in the horizontal axis on a given survey screen
log_aY	numeric	natural logarithm of the average absolute cursor acceleration (in pixels per second squared) in the vertical axis on a given survey screen

Table A5.2

Auxiliary Information for Files Describing Separate Respondents

Variable	Format	Description
idResp	integer	respondent’s reshuffled ID
token	string	respondent’s ID
study	integer: {0, 1}	in which part of the study the test was taken: 0 – main, 1 – pilot
status	integer: {0, 1}	quality control (QC) status of the test: 0 – passed QC, 1 – failed QC
badItem	integer: {0, 1}	which version of the “bad” item was answered by a given test-taker: 0 – version with no defect, 1 – version with a fault (i.e. many correct answers)
itemSetsOrder	integer: {0, 1}	order of presentation of testlets during the test: 0 – testlet one first, 1 – testlet two first
instructionOrder	integer: {0, 1}	order of instruction during the test: 0 – attentive instruction first, 1 – inattentive instruction first
localitySize	integer: {0, 1, 2}	school location size: 0 – village, 1 – small town, 2 – big town/city
grade	integer: {0, 1}	test-taker’s grade: 0 – seventh, 1 – eighth
gender	integer: {0, 1, 2}	test-taker’s gender: 0 – male, 1 – female
motivation	integer: {0, 1, 2, 3}	test-taker’s self declared motivation (“How hard did you try during the test?”): 0 – Not at all, 1 – Not much, 2 –Rather hard, 3 – Very hard
strategy2	integer: {0, 1, 2, 3, 4}	test-taker’s self declared response strategy under the inattentive instruction: 0 – solving only easy items, 1 – informed guessing, 2 – attentive answering (indicates a failure of manipulation), 3 – speeding, 4 – other
books	integer: {0, 1, 2, 3, 4}	number of books in the test-taker’s home (self-declared): 0 – 0-10, 1 – 11-25, 2 – 26-100, 3 – 101-200, 4 – >200
mouseUse	integer: {0, 1, 2, 3}	test taker’s self declared way of using a mouse: 0 – “tracking what I read”, 1 – “marking answers only”, 2 – other “unrelated moves” 3 - “missing answer”
scoreAtt	numeric	test-taker’s sum score of the testlet with attentive instruction
scoreInatt	numeric	test-taker’s sum score of the testlet with inattentive instruction
M032579_answer	integer: {0, 1, 2, 3} | missing	answer on item M032579: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D

Variable	Format	Description
M042254_answer	integer: {0, 1, 2, 3} | missing	answer on item M042254: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M042222_answer	integer: {0, 1, 2, 3} | missing	answer on item M042222: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gm2006z08_answer	integer: {0, 1, 2, 3} | missing	answer on item gm2006z08: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M042260_answer	integer: {0, 1, 2, 3} | missing	answer on item M042260: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gmm2015z07_answer	integer: {0, 1, 2, 3} | missing	answer on item gmm2015z07: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M032295_answer	integer: {0, 1, 2, 3} | missing	answer on item M032295: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M042031_answer	integer: {0, 1, 2, 3} | missing	answer on item M042031: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gm2004z04_answer	integer: {0, 1, 2, 3} | missing	answer on item gm2004z04: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M032679_answer	integer: {0, 1, 2, 3} | missing	answer on item M032679: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M032198_answer	integer: {0, 1, 2, 3} | missing	answer on item M032198: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gmm2014z11_answer	integer: {0, 1, 2, 3} | missing	answer on item gmm2014z11: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M042041_answer	integer: {0, 1, 2, 3} | missing	answer on item M042041: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
e82022z7_answer	integer: {0, 1, 2, 3, 4} | missing	answer on item e82022z7: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D, 4 – answer E
M032166_answer	integer: {0, 1, 2, 3} | missing	answer on item M032166: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D

Variable	Format	Description
M042032_answer	integer: {0, 1, 2, 3} | missing	answer on item M042032: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M052302_answer	integer: {0, 1, 2, 3} | missing	answer on item M052302: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gmm2016z20_answer	integer: {0, 1, 2, 3} | missing	answer on item gmm2016z20: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M032738_answer	integer: {0, 1, 2, 3} | missing	answer on item M032738: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M032477_answer	integer: {0, 1, 2, 3} | missing	answer on item M032477: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gmm2018z07_answer	integer: {0, 1, 2, 3} | missing	answer on item gmm2018z07: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M032704_answer	integer: {0, 1, 2, 3} | missing	answer on item M032704: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M032116_answer	integer: {0, 1, 2, 3} | missing	answer on item M032116: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gmm2016z01_answer	integer: {0, 1, 2, 3} | missing	answer on item gmm2016z01: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M042236_answer	integer: {0, 1, 2, 3} | missing	answer on item M042236: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gm2002z05_answer	integer: {0, 1, 2, 3} | missing	answer on item gm2002z05: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
gmm2016z03_answer	integer: {0, 1, 2, 3} | missing	answer on item gmm2016z03: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M022101_answer	integer: {0, 1, 2, 3} | missing	answer on item M022101: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
M052084_answer	integer: {0, 1, 2, 3} | missing	answer on item M052084: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D

Variable	Format	Description
M032094_answer	integer: {0, 1, 2, 3} | missing	answer on item M032094: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D
badBad_answer	integer: {0, 1, 2, 3, 4} | missing	score on the “bad” item – faulty version: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D, 4 – answer E
badGood_answer	integer: {0, 1, 2, 3, 4} | missing	score on the “bad” item – version with no defect: 0 – answer A, 1 – answer B, 2 – answer C, 3 – answer D, 4 – answer E
M032579_score	integer: {0, 1}	score on item M032579: 0 – incorrect answer or missing answer, 1 – correct answer
M042254_score	integer: {0, 1}	score on item M042254: 0 – incorrect answer or missing answer, 1 – correct answer
M042222_score	integer: {0, 1}	score on item M042222: 0 – incorrect answer or missing answer, 1 – correct answer
gm2006z08_score	integer: {0, 1}	score on item gm2006z08: 0 – incorrect answer or missing answer, 1 – correct answer
M042260_score	integer: {0, 1}	score on item M042260: 0 – incorrect answer or missing answer, 1 – correct answer
gmm2015z07_score	integer: {0, 1}	score on item gmm2015z07: 0 – incorrect answer or missing answer, 1 – correct answer
M032295_score	integer: {0, 1}	score on item M032295: 0 – incorrect answer or missing answer, 1 – correct answer
M042031_score	integer: {0, 1}	score on item M042031: 0 – incorrect answer or missing answer, 1 – correct answer
gm2004z04_score	integer: {0, 1}	score on item gm2004z04: 0 – incorrect answer or missing answer, 1 – correct answer
M032679_score	integer: {0, 1}	score on item M032679: 0 – incorrect answer or missing answer, 1 – correct answer
M032198_score	integer: {0, 1}	score on item M032198: 0 – incorrect answer or missing answer, 1 – correct answer
gmm2014z11_score	integer: {0, 1}	score on item gmm2014z11: 0 – incorrect answer or missing answer, 1 – correct answer
M042041_score	integer: {0, 1}	score on item M042041: 0 – incorrect answer or missing answer, 1 – correct answer
e82022z7_score	integer: {0, 1}	score on item e82022z7: 0 – incorrect answer or missing answer, 1 – correct answer
M032166_score	integer: {0, 1}	score on item M032166: 0 – incorrect answer or missing answer, 1 – correct answer
M042032_score	integer: {0, 1}	score on item M042032: 0 – incorrect answer or missing answer, 1 – correct answer

Variable	Format	Description
M052302_score	integer: {0, 1}	score on item M052302: 0 – incorrect answer or missing answer, 1 – correct answer
gmm2016z20_score	integer: {0, 1}	score on item gmm2016z20: 0 – incorrect answer or missing answer, 1 – correct answer
M032738_score	integer: {0, 1}	score on item M032738: 0 – incorrect answer or missing answer, 1 – correct answer
M032477_score	integer: {0, 1}	score on item M032477: 0 – incorrect answer or missing answer, 1 – correct answer
gmm2018z07_score	integer: {0, 1}	score on item gmm2018z07: 0 – incorrect answer or missing answer, 1 – correct answer
M032704_score	integer: {0, 1}	score on item M032704: 0 – incorrect answer or missing answer, 1 – correct answer
M032116_score	integer: {0, 1}	score on item M032116: 0 – incorrect answer or missing answer, 1 – correct answer
gmm2016z01_score	integer: {0, 1}	score on item gmm2016z01: 0 – incorrect answer or missing answer, 1 – correct answer
M042236_score	integer: {0, 1}	score on item M042236: 0 – incorrect answer or missing answer, 1 – correct answer
gm2002z05_score	integer: {0, 1}	score on item gm2002z05: 0 – incorrect answer or missing answer, 1 – correct answer
gmm2016z03_score	integer: {0, 1}	score on item gmm2016z03: 0 – incorrect answer or missing answer, 1 – correct answer
M022101_score	integer: {0, 1}	score on item M022101: 0 – incorrect answer or missing answer, 1 – correct answer
M052084_score	integer: {0, 1}	score on item M052084: 0 – incorrect answer or missing answer, 1 – correct answer
M032094_score	integer: {0, 1}	score on item M032094: 0 – incorrect answer or missing answer, 1 – correct answer
badBad_score	integer: {0, 1} | missing	score on the “bad” item – fault version: 0 – incorrect answer or missing answer, 1 – correct answer
badGood_score	integer: {0, 1} | missing	score on the “bad” item – version with no defect: 0 – incorrect answer or missing answer, 1 – correct answer
M032579GroupTime	numeric: >0	time spent on item M032579 [seconds]
M042254GroupTime	numeric: >0	time spent on item M042254 [seconds]
M042222GroupTime	numeric: >0	time spent on item M042222 [seconds]
gm2006z08GroupTime	numeric: >0	time spent on item gm2006z08 [seconds]
M042260GroupTime	numeric: >0	time spent on item M042260 [seconds]

Variable	Format	Description
gmm2015z07GroupTime	numeric: >0	time spent on item gmm2015z07 [seconds]
M032295GroupTime	numeric: >0	time spent on item M032295 [seconds]
M042031GroupTime	numeric: >0	time spent on item M042031 [seconds]
gm2004z04GroupTime	numeric: >0	time spent on item gm2004z04 [seconds]
M032679GroupTime	numeric: >0	time spent on item M032679 [seconds]
M032198GroupTime	numeric: >0	time spent on item M032198 [seconds]
gmm2014z11GroupTime	numeric: >0	time spent on item gmm2014z11 [seconds]
M042041GroupTime	numeric: >0	time spent on item M042041 [seconds]
e82022z7GroupTime	numeric: >0	time spent on item e82022z7 [seconds]
M032166GroupTime	numeric: >0	time spent on item M032166 [seconds]
M042032GroupTime	numeric: >0	time spent on item M042032 [seconds]
M052302GroupTime	numeric: >0	time spent on item M052302 [seconds]
gmm2016z20GroupTime	numeric: >0	time spent on item gmm2016z20 [seconds]
M032738GroupTime	numeric: >0	time spent on item M032738 [seconds]
M032477GroupTime	numeric: >0	time spent on item M032477 [seconds]
gmm2018z07GroupTime	numeric: >0	time spent on item gmm2018z07 [seconds]
M032704GroupTime	numeric: >0	time spent on item M032704 [seconds]
M032116GroupTime	numeric: >0	time spent on item M032116 [seconds]
gmm2016z01GroupTime	numeric: >0	time spent on item gmm2016z01 [seconds]
M042236GroupTime	numeric: >0	time spent on item M042236 [seconds]
gm2002z05GroupTime	numeric: >0	time spent on item gm2002z05 [seconds]
gmm2016z03GroupTime	numeric: >0	time spent on item gmm2016z03 [seconds]
M022101GroupTime	numeric: >0	time spent on item M022101 [seconds]
M052084GroupTime	numeric: >0	time spent on item M052084 [seconds]
M032094GroupTime	numeric: >0	time spent on item M032094 [seconds]
badItemGroupTime	numeric: >0	time spent on the “bad” item [seconds]
sumItemsScreenTimeAtt	numeric: >0	total time spent on testlet with attentive instruction [seconds]
sumItemsScreenTimeInatt	numeric: >0	total time spent on testlet with inattentive instruction [seconds]
sumItemsScreenTimeLogAtt	numeric	natural logarithm of the total time spent on testlet with attentive instruction [seconds]
sumItemsScreenTimeLogInatt	numeric	natural logarithm of the total time spent on testlet with inattentive instruction [seconds]

Variable	Format	Description
scoreAttTercile	integer: {0, 1, 2}	tercile of test-taker's sum score of the testlet with attentive instruction
scoreInattTercile	integer: {0, 1, 2}	tercile of test-taker's sum score of the testlet with inattentive instruction
mahalanobisAtt	numeric: >0	mahalanobis distance of test-taker's score vector in testlet with attentive instruction
mahalanobisInatt	numeric: >0	mahalanobis distance of test-taker's score vector in testlet with inattentive instruction
log_dXAtt	numeric	natural logarithm of the total distance (in pixels) travelled by the cursor in the horizontal direction on a given survey screen – testlet with attentive instruction
log_dXInatt	numeric	natural logarithm of the total distance (in pixels) travelled by the cursor in the horizontal direction on a given survey screen – testlet with inattentive instruction
log_dYAtt	numeric	square root of the number of flips (changes in the cursor move direction) in the horizontal axis on a given survey screen – testlet with attentive instruction
log_dYInatt	numeric	square root of the number of flips (changes in the cursor move direction) in the horizontal axis on a given survey screen – testlet with inattentive instruction
sqrt_flipsXAtt	numeric: >0	natural logarithm of the total distance (in pixels) travelled by the cursor in the vertical direction on a given survey screen – testlet with attentive instruction
sqrt_flipsXInatt	numeric: >0	natural logarithm of the total distance (in pixels) travelled by the cursor in the vertical direction on a given survey screen – testlet with inattentive instruction
sqrt_flipsYAtt	numeric: >0	square root of the number of flips (changed of the cursor move direction) in the vertical axis on a given survey screen – testlet with attentive instruction
sqrt_flipsYInatt	numeric: >0	square root of the number of flips (changed of the cursor move direction) in the vertical axis on a given survey screen – testlet with inattentive instruction
log_vXAtt	numeric	natural logarithm of the average cursor speed (in pixels per second) in the horizontal axis on a given survey screen – testlet with attentive instruction

Variable	Format	Description
log_vXInatt	numeric	natural logarithm of the average cursor speed (in pixels per second) in the horizontal axis on a given survey screen – testlet with inattentive instruction
log_aXAtt	numeric	natural logarithm of the average cursor speed (in pixels per second) in the vertical axis on a given survey screen – testlet with attentive instruction
log_aXInatt	numeric	natural logarithm of the average cursor speed (in pixels per second) in the vertical axis on a given survey screen – testlet with inattentive instruction
log_vYAtt	numeric	natural logarithm of the average absolute cursor acceleration (in pixels per second squared) in the horizontal axis on a given survey screen – testlet with attentive instruction
log_vYInatt	numeric	natural logarithm of the average absolute cursor acceleration (in pixels per second squared) in the horizontal axis on a given survey screen – testlet with inattentive instruction
log_aYAtt	numeric	natural logarithm of the average absolute cursor acceleration (in pixels per second squared) in the vertical axis on a given survey screen – testlet with attentive instruction
log_aYInatt	numeric	natural logarithm of the average absolute cursor acceleration (in pixels per second squared) in the vertical axis on a given survey screen – testlet with inattentive instruction