

Final Report for the IEA R&D 2023

Research Project entitled:

Using Large Language Models for Automatic Item Generation: Development and Validation for TIMSS Grade 4

PI: Marek Muszyński

Co-Is: Hubert Plisiecki, Artur Pokropek, Tomasz Żółtak

Institute of Philosophy and Sociology of the Polish Academy of Sciences
(IFiS PAN)

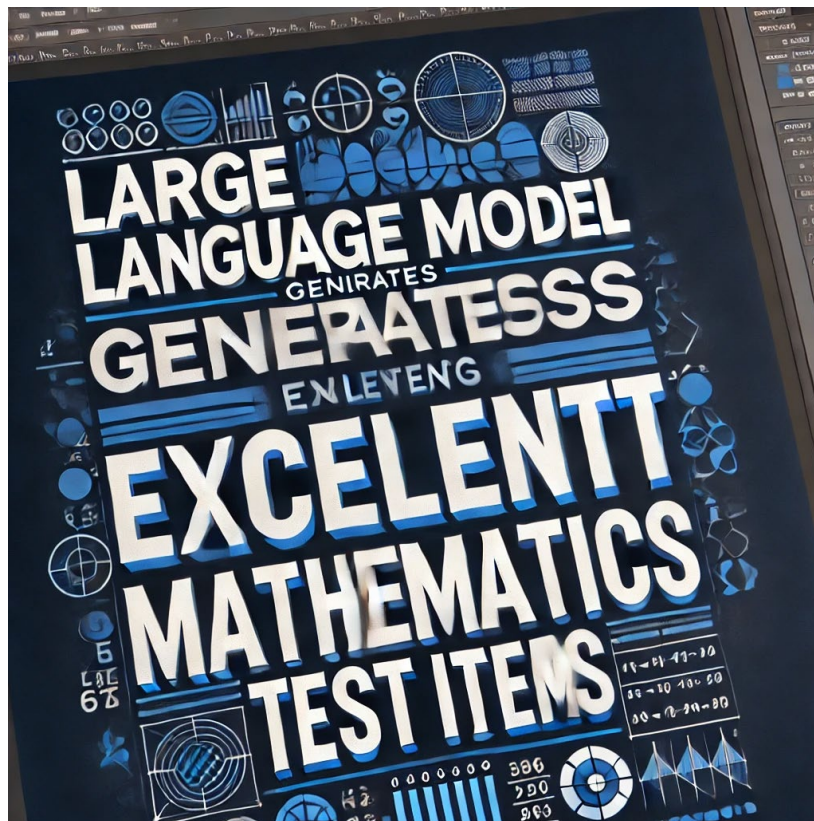


Table of contents

Executive Summary	4
1. Introduction and study aims	5
2. Study preparation	8
2.1 Item extraction	8
2.2 Item Generation Design	9
2.3 Item generation and initial revision process	15
3. Subject matter experts panel	16
3.1 Subject matter experts panel design	16
3.2 Subject matter expert panel composition	19
3.3 SMEs panel results	22
3.3.1 Agreement Analyses	22
3.3.2 Raters' evaluations	26
3.3.2.1 Mathematics	26
3.3.2.2 Science	34
3.3.2.3 Additional Analyses	42
3.3.3 Raters' comments	43
3.4 SMEs Panel Summary and Discussion	47
3.4.1 Limitations	48
3.4.2 Future directions	49
3.4.3 Conclusions	49
4. Validation study	50
4.1 Item generation	50
4.1.1 Overall Approach	50
4.1.2 The Simple Setup	50
4.1.3 The Complex Setup	58
4.1.4 The Final Choice of Items	59
4.1.5 Generating Images	60
4.2 Test preparation	60
4.2.1 Test plan (incomplete block design)	60
4.2.2 Assessment contents	62
4.2.3 Paradata collection	64
4.3 Fieldwork preparation, procedures and timeline	64
4.3.1 School recruitment and remuneration	64
4.3.2 Testing room preparation	65
4.3.3 Assessment session procedures	66
4.3.4 Pilot administration	68
4.3.5 Data collection timeline	68
4.3.6 Sample composition	72
4.3.7 Booklet Coverage	73

4.3.8 Item coding	74
4.4 Psychometric modeling	77
4.4.1 Excluded items	77
4.4.2 Model estimation and model fit	78
4.4.3 Item parameters of TIMSS and LLM-generated items	82
4.4.4 Differential Item Functioning	102
4.4.5 Testing Local Item Independence	105
4.4.6 Summary of psychometric analyses	107
5. General discussion	112
5.1 Results summary and conclusion	112
5.2 Future directions	115
6. References	120
7. Supplement	125
7.1 Duplicates removed from the generated item set	125
7.2 Kruskal-Wallis Test Results	128
7.3 Additional results for DIF analysis	134
7.4 List of Annexes	144

Executive Summary

The study aimed to validate the quality of assessment items generated by Large Language Models for use in mathematics and science assessment on the example of the TIMSS Grade 4. The validation process included expert ratings and qualitative assessment, as well as using the generated items in a field test that enabled the assessment of the substantial and psychometric properties of the generated items. Since LLM-generated items were mixed with original TIMSS items, we were able to compare the two item types directly.

The study's objective was to assess the current possibilities of the widely accessible models without a heavy reliance on human editing and selecting. To this end, we did not edit the generated items nor selected them based on any criteria apart from correcting minor and obvious errors (e.g., spelling) and removing repeated items.

The gathered empirical evidence showed that many of the LLM-generated items were of high quality. However, overall, the subject matter experts rated the generated items as less suitable for assessments and less attractive compared to original TIMSS items. Psychometrically, the generated items also performed less well than real TIMSS items. The main psychometric issues included: difficulty level not aligned with the target group, presence of items with negative or low discrimination parameter, and violations of the local independence assumption. No differences were found between TIMSS and generated items in terms of differential item functioning, item fit, or distractor functionality.

The study also analysed the most common problems observed in the generated items and their underlying causes. The generated items were often too easy (especially the multiple-choice mathematics items) or too difficult (especially the constructed-response science items). LLMs frequently produced very similar and repetitive items. The response options generated also displayed several issues, such as implausible or unjustified distractors, repeated distractors, absence of a correct answer, or the presence of multiple correct answers.

The LLM also demonstrated difficulty in using precise terminology and appropriate style. Occasional problems with style and grammar may have stemmed from the use of Polish as the assessment language, which has relatively less digital content available for LLM training compared to English (a common issue for “resource-poor” languages). Overall, the mathematics items appeared to perform better than the science items.

We recommend continuing this line of research using domain-specific fine-tuned models run locally, as they currently hold the most promise for producing high-quality LLM-generated items. Human intervention and judgment remain crucial for selecting the best items.

1. Introduction and study aims

Items constitute a fundamental aspect of any assessment. A comprehensive measurement relies on psychometrically sound, linguistically correct, and theoretically aligned items. The growing demand for measurement programs resulted in an increased need for such items. However, developing excellent items requires daunting amounts of precious resources. Typically, it involves teams of subject matter experts, psychometricians, and fieldworkers, who are responsible for developing, evaluating, piloting, and validating items before they enter fieldwork. The process is long, costly, and labor-intensive.

For a long time, researchers aimed at streamlining the process of item development. This line of research was named automatic item generation (AIG) and uses different methods to shortcut the process of item “production”. First, means such as item cloning were used (e.g., Glas & van der Linden, 2003). Here, very close duplicates of a model item are generated with essentially the same content but small changes to the item’s stem and/or distractors. This framework was later developed to generate item clones on the basis of predefined item models and computer algorithms that helped to alter key item elements to get new items (Gierl et al., 2016). The number and quality of these “automatically” generated items depended on the number and quality of prepared item models and item elements that were set to vary between item versions. The whole process was rather semirather than fully automatic and required a lot of human work (Attali et al., 2022).

Next, approaches based on deep learning started to emerge, where computer algorithms specialized in natural language processing were utilized to create elements of new items (stems, vignettes, distractors, etc.). At this stage, neural networks were mostly employed for AIG, the process remained long and difficult, as early models needed time-consuming and often costly pre-training/fine-tuning that required significant technical knowledge. The outcomes were often far from perfect, as the created items were very basic and necessitated further human polishing up. Examples of these early works are Gao et al. (2019), Liang et al. (2018), Shin et al. (2019). Susanti et al. (2018) and Zu et al. (2023) investigated methods of automatic distractor creation for item stems. Mitkov et

al. (2023) showed a method of creating whole items (stem + distractors) based on text passages used to train neural networks. Von Davier (2023) and Bezirhan and von Davier (2023) presented a method to create text passages for further item creation based on GPT-2 and GPT-3 models (early OpenAI models, see cited papers for more details).

The rapid development of large language models (LLMs) lead to the emergence of new tools that are now capable of excellent processing and generating of natural language texts and other tasks (e.g. Bubeck et al., 2023). The increase in abilities of these models calls for a revisit of previous research on this topic. Recently, Drori et al. (2022) or Attali et al. (2022) came up with methods to use large language models to construct whole items in the domain of mathematics and English, respectively. Several studies succeeded in automatically developing self-report items, e.g. personality scales (Götz et al., 2023; Hommel et al. 2022; von Davier, 2018). What is more, the initial research shows that it is possible to automatically generate items of pre-set difficulty and complexity (Raina & Gales, 2022; Settles et al., 2020). For example, Bezirhan and von Davier (2023) demonstrated that LLMs require only short prompts to generate items of varying difficulty (as assessed by the experts).

Another research area opened by the use of LLMs is prompt engineering, research on developing and optimizing prompts to use models efficiently (Reynolds & McDonell, 2021). The outcome of AIG differs depending on the quantity and quality of the context used to prompt LLMs and the exact question (instruction) posed. Wang et al. (2022) compared the items generated with the use of prompts differing in number and complexity of examples, as well as length of the overall context provided. In this study, items of higher quality were generated with more content-specific examples, with five, being the optimal number of examples provided. Shorter or medium-length context information (15-25 word-long) yielded generation of “better” items than long context (40 words; Wang et al., 2022). On the other hand, zero-shot prompts (no examples) were as effective as one-shot prompts (one example) in the context of GPT-3 and creating PIRLS passages (Bezirhan & von Davier, 2023).

The research on modelling task difficulty by prompts and most effective way of prompting is just at its beginning (Attali et al., 2022). However, it is already evidenced that LLMs can incorporate (“learn”) knowledge provided “as contextual information within

prompts” (Lampinen et al., 2022; Wang, 2023) as an alternative to fine-tuning them with an updated dataset (Drori et al., 2022). Prompting is naturally more time- and cost-effective, especially that newer LLMs allow for longer and more complex prompts with a lot of context provided.

This project aims to broaden the findings of early studies and show that AIG can be employed in the rigorous context of international large-scale assessments, such as TIMSS. Moreover, we intend to go further than previous studies in the context of IEA’s assessments, e.g. Bezirhan and von Davier (2023) presented a method to create text passages for further creation of PIRLS Grade 4 items and got the created passages evaluated by human raters. Here, we aim to generate complete items (stem and response options) and then subject them to human experts’ evaluation but also to validate them in a field test mimicking real IEA assessment setting. We will also focus on developing items closely aligned to the TIMSS assessment framework, as well as generating items of pre-defined difficulty. Graphical item elements, common for TIMSS Grade 4, will be generated by graphical generative artificial intelligence programs, such as DALL-E 3 or with the use of graphical plug-ins for GPT-4. Recently, GPT-4 was equipped with new plug-ins that link with Python to produce tables and charts, which should be very handy for generating TIMSS 4 items. In short, the main research items are:

1. Develop methods of automatic item generation (AIG) using large language models (LLMs) for maths & science assessment TIMSS Grade 4 (in Polish).
2. Validate these methods using a subject matter experts panel and a school-based assessment following closely the original TIMSS Grade 4 methodology and design.
3. Provide evidence-based knowledge on the possibilities of employing AIG for international large-scale assessments based on the examples generated for TIMSS Grade 4. This is a fast developing domain and as such the results of the study (Summer 2024) might not be fully generalizable to the current (September 2025) capabilities of LLM models.

2. Study preparation

2.1 Item extraction

TIMSS Grade 4 restricted use items in English and Polish were obtained from the IEA. The documents with Polish TIMSS items were passed to the OpenAI's "gpt4-o" image API, page after page, with the following query:

Extract exercises from this TIMSS exam page image in a json format.

Whenever tables appear, encode them in the question variable using markdown alongside the exercise text. Whenever images appear, describe what they portray in Polish in the format [OBRAZEK: image_description] and add the description to the question variable.

The important variables are:

"exercise_id" (example formats: M_01_07; S_02_05)

"question" : exercise text

"answers" : answers

"contains_image" (boolean)

Make sure to produce valid jsons, escape characters properly, extract all exercises, and do not write anything else!

This allowed us to extract 174 TIMSS science items TIMSS, and 177 TIMSS maths items (both Grade 4). The extracted items were then manually checked for hallucinations, and formatting problems by the members of the research team. While checking the items, the research team members also supplemented the database with the international IDs of the items by adding this argument to every item in the json.

The metadata related to content, and cognitive domain for the items was manually copied from the relevant TIMSS documents.

A short comment applies to data safety and research reproducibility. According to the company's policy (<https://openai.com/api/>), the API does not use the data to train AI models and does not process it in any other kind. Moreover, the data is removed from the platform automatically after an indicated amount of time

(<https://platform.openai.com/docs/guides/your-data>). The OpenAI API use was consulted with the IEA before any operations took place. Hence, using the API, along with other “business” options to access OpenAI models seem safe from the data protection point of view.

With regard to the research reproducibility and open science, we cannot provide the IEA restricted use items employed to prompt the models, but any interested researcher can apply to use them for research purposes. Having IEA’s consent and our prompts, anyone can reproduce the process. However, we cannot guarantee full reproduction, as the OpenAI models constantly change - and we can do nothing with that. This issue will be raised again in the discussion.

2.2 Item Generation Design

For the pilot study, the items were generated using the “GTP-4o” API from OpenAI. For the hyperparameters, the only non-default parameter was the temperature¹, which was set to 0.8 based on a short generation pilot study, where we iteratively generated items until we have finetuned this parameter to allow for maximum variability, without perceived negative impact on item intelligibility. Testing this and other parameters in a systematic manner was decided against due to the time and resource consuming iterative process that it would have required. The number of completions (n) was left at its default of 1, and both `presence_penalty` and `frequency_penalty` remained at 0, so no additional bias for novelty or against repetition was imposed on the outputs. Because `max_tokens` was not set, the model was free to generate up to its full 4 096-token completion limit within GPT-4o’s 128 000-token context window, effectively letting the answer run until the model decided to stop. Different multiple-shot setups were used, where different numbers of example TIMSS items were submitted to the LLM alongside the request to generate new items.

¹ Temperature is the parameter to adjust the outputs of large language models, the higher the temperature, the higher the randomness and more diverse the outputs. However, the parameter cannot be directly linked to “creativity” (Peeperkorn et al., 2024).

The setups were:

- zero-shot - 0 examples,
- one-shot - 1 example,
- three-shot - 3 examples,
- and five-shot - 5 examples.

The choice of 0, 1, 3, and 5 examples was made to systematically explore model behavior under different levels of supervision, in a way that is appropriate for a new task where no established few-shot benchmarks yet exist.

Given the novelty of the task, it was important to treat the evaluation setup exploratorily rather than relying on conventions that may have been developed for substantially different problems. Predefined standards from other tasks may not necessarily translate well to this context, and could introduce unjustified assumptions about what levels of supervision are informative. We therefore adopted a progression of 0, 1, 3, and 5 examples.

Our choice was also based on literature. Laverghetta Jr. & Licato (2023) used three examples, while Drori et al. (2022) used a wider array of examples, starting from zero-shot and ending with up to five examples. Wang et al. (2022) used an even larger number of examples, reaching as many as seven, but concluded that five examples led to best results. Other studies used fewer examples, e.g. Bezirhan & von Davier (2023) - zero and one, Omopekunola & Yu Kardanova (2024) - from one to three. Hence, our prompting scheme used the most common number of examples used in the literature.

This incremental design ensures that we can capture meaningful differences in model behavior as limited supervision is introduced, while avoiding the risks of overfitting to an arbitrary setup. In the absence of preexisting standards for this task, we believe this approach is both reasonable and scientifically justified.

For each setup, 80 items were generated by the LLM—40 in science and 40 in mathematics. Each query specified both the content domain and the cognitive domain, which were randomly drawn based on the metadata of real TIMSS items. This ensured that the distribution of domains among the generated items reflected the actual TIMSS

proportions. Example items were drawn from the TIMSS dataset, limited to those matching the selected content and cognitive domains.

The prompt used to generate the items was created in an iterative manner, where the members of the team inspected some example-generated items and refined the prompt. Its final form is presented below; please mind that the words in curly brackets were automatically swapped with the specific values chosen per specific generation instance:

```
"Create one item for the Polish TIMSS assessment. Do not write anything else than the content of the item. The item has to be in Polish.  
Subject : {subject}  
Content Domain : {content_domain}  
Definition: {content_domain_definitions[content_domain]}  
Cognitive Domain : {cognitive_domain}  
Definition: {cognitive_domain_definitions[cognitive_domain]}  
Age appropriateness: Grade 4 of the elementary school  
Item format: Constructed response OR multiple-choice. Constructed response items usually require students to give a numerical result, provide a short explanation or description given in one or two phrases or sentences, create a list or complete a table. Multiple choice items have four response options, ONE OF THEM IS ALWAYS CORRECT AND THREE ARE INCORRECTt. The incorrect options need to be plausible choices, demonstrating typical errors or misunderstandings. They need to be similar to the correct option in terms of length and format, but be undoubtedly incorrect.  
Additional comments: DO NOT GENERATE LINKS TO TABLES AND IMAGES, BUT PRESENT A DETAILED DESCRIPTION OF TABLES AND IMAGES IF YOU CHOOSE TO USE THEM IN YOUR ITEM (YOU DONT HAVE TO)  
"
```

The prompt included the definitions of the sampled content and cognitive domain. These were prepared on the basis of the TIMSS assessment framework (Mullis et al., 2023, pp. 7-10 & 14-18 for maths; pp. 20-28 & 40-44) and are shown below:

Cognitive Domains:

Knowing

Facility in applying mathematics, or reasoning about mathematical situations, depends on familiarity with mathematical concepts and fluency in mathematical skills. The more relevant knowledge a student is able to recall and the wider the range of concepts he or she understands, the greater the potential for engaging with a wide range of problem situations. It contains skills like Recall: Recall definitions, terminology, number properties, units of measurement, geometric properties, and notation (e.g., $a \times b = ab$, $a + a + a = 3a$). Identify: Identify numbers, expressions, quantities, and shapes. Recognize when entities are mathematically equivalent. Read information from graphs, tables, texts, or other sources. Order: Order and classify numbers, expressions, quantities, and shapes by common properties. Compute: Compute arithmetic operations with whole numbers, fractions, decimals, and integers using algorithmic procedures. Carry out straightforward algebraic manipulation.

Applying

The applying domain involves the application of mathematics in a range of situations. Problem-solving is central to this domain. Students will need to select suitable operations, strategies, and tools for solving problems. Formulate: Determine efficient/appropriate operations, strategies, and tools for solving problems. Implement: Implement suitable strategies and operations to produce solutions to problems. Represent: Represent data in tables or graphs; create equations, inequalities, geometric figures, or diagrams that model problem situations; and generate equivalent representations for a given mathematical entity or relationship.

Reasoning

Reasoning mathematically involves logical, systematic thinking. It includes intuitive and inductive reasoning based on patterns and regularities that can be used to arrive at solutions to problems. Evidence of reasoning processes can be found in the explaining

or justifying of a solution method, or the making of valid inferences on the basis of information and evidence. Reasoning is required in analyzing or generalizing mathematical relationships. It contains skills like: Analyze: Analyze, describe, or use relationships among numbers, expressions, quantities, and shapes. Integrate: Link different elements of knowledge, related representations, and procedures. Generalize: Make statements that represent relationships in more general and more widely applicable terms. Justify: Provide mathematical arguments to support a strategy or solution.

Content Domains

Number

Number provides the foundation of mathematics in primary school. The number content domain consists of three topic areas: Whole numbers, Expressions, simple equations and relationships, and Fractions and decimals.

Geometric Shapes and Measures

Geometry helps us visualize and understand the relationships between shapes and sizes. Measurement is the process of quantifying attributes of objects and phenomena (e.g., length and time). At the fourth grade, students should be able to use a ruler to measure length; perform calculations involving length, mass, volume, and time; calculate areas of shapes based on rectangles; calculate perimeters of polygons; and use cubes to determine volumes. Students should be able to identify the properties and characteristics of lines, angles, and a variety of two- and three-dimensional shapes. Spatial sense is integral to the study of geometry, and students will be asked to describe and draw a variety of geometric figures. They also should be able to analyze geometric relationships and use these relationships to draw conclusions about geometric objects.

Data Display

Students need to understand that graphs and charts help organize information or categories and provide a way to compare data. The data content domain consists of two topic areas: Reading and displaying data, Interpreting, combining, and comparing data. At the fourth grade, students should be able to read and create data displays. They should be able to make inferences from data displays and use data from one or more sources to answer questions of interest.

Life Science

Life science is represented by five topic areas: Characteristics and life processes of organisms, Life cycles, reproduction, and heredity, Organisms, environment, and their interactions, Ecosystems, and Human health.

Physical Science

Physical science content domain comprises: Classification and properties of matter and changes in matter, Forms of energy and energy transfer, and Forces and motion. Fourth grade students should have an understanding of physical states of matter (solid, liquid, and gas), as well as common changes in the state and form of matter; this forms a foundation for the study of both chemistry and physics in the middle and upper grades. At this level, students also should know common forms and sources of energy and their practical uses, and understand basic concepts about light, sound, electricity, and magnetism. The study of forces and motion emphasizes an understanding of forces as they relate to movements students can observe, such as the effect of gravity or pushing and pulling.

Earth Science

Earth science is the study of Earth and its place in the Solar System, and at fourth grade focuses on the study of phenomena and processes that students can observe in their everyday lives. Topic areas included in this domain are generally considered to be important for students at the fourth grade to understand as they learn about the planet on which they live and its place in the Solar System: Earth's physical characteristics, resources, and history, Earth's weather and climates, and Earth in the Solar System. At this level, students should have some general knowledge about the structure and physical characteristics of Earth's surface, and about the use of Earth's most important resources. Students also should be able to describe some of Earth's processes in terms of observable changes and understand the time frame over which such changes have occurred. Fourth grade students should also demonstrate some understanding about Earth's place in the Solar System based on observations of patterns of change on Earth and in the sky.

2.3 Item generation and initial revision process

As can be seen above, the cognitive domain definitions were related only to mathematics, which was a technical mistake in the code preparation. This will be taken into consideration when analysing the results of the quality inspection of the items.

The first 20 items were chosen from each subject/setup cross-section in each domain (math and science). To avoid too much human interference, we just selected the first 20 items from each of the setups, discarding only the items clearly unusable, which were replaced by next items. Nine items had to be dropped as they were unusable in an obvious manner. Those cases were mostly duplicates (repetitive items), four of them came from the zero-shot condition, four from the one-shot condition and one from the five-shot condition. We present details on the process of removing duplicates in Supplement 1.

The generated items were then copied to an MS Excel spreadsheet and for those that required an image, the research team members created the images using the web based "GPT-4" model. The creation of images was an interactive process, and required

the team members to prompt the model with regard to what kind of image it had to create based on the content of the item. The research members had to choose and prompt the model to either use the DALL-E 3 capabilities to generate the images or to use LLM-generated Python and Matplotlib code to generate the diagrams and plots. DALL-E 3 is a state-of-the-art diffusion-based text-to-image model that creates full-resolution pictures from natural-language prompts, whereas Matplotlib is a mature Python library that programmatically renders vector graphics and charts, giving the author pixel-perfect control over every line, marker, and axis. The choice was driven by whether the image required precisely drawn geometric shapes (e.g., specific math items) or whether it could be presented illustratively (e.g., an image of a specific number of planets). Therefore, whenever an image could be drawn as geometric shapes, Matplotlib prompting was used. This was done because images created using the DALL-E 3 model were highly nondeterministic — it was hard to specify details such as the precise number of objects in the image or their exact shapes, as DALL-E 3 would frequently reinterpret the prompt in stylistic ways that broke those strict constraints. Much of this generation process was therefore left at the hands of the researchers, relying on their intuition and revision process to generate the best images. After the images were created, the text of the items was cleaned and formatted.

The final spreadsheet consisted of 100 items in each domain: 20 items from each of the generating setups (zero-, one-, three-, and five-shot) to a total of 80 LLM-generated items and 20 original TIMSS items.

3. Subject matter experts panel

3.1 Subject matter experts panel design

The first part of the instructions described the structure and handling of the Evaluation Sheet, which, for assessing mathematics test tasks, was structured as follows:

- General Layout: The sheet consists of rows and columns where tasks are listed from row 2 to row 101. Each row represents a separate task.
- Content Columns:

- Column "B" displays the task description.
- Columns "C" and "D" may include graphical elements such as images, diagrams, or charts associated with the tasks.
- Columns "E" to "H" can contain graphic options for answers.
- Evaluation Criteria:
 - The criteria for evaluating tasks are located in columns "I" to "S".
 - Each criterion is in a separate column, with the first row containing the column names and criteria descriptions, which remain visible during scrolling.
- Navigation: Scroll down using the scrollbar on the right side of the screen or the mouse wheel to move to subsequent tasks. Scroll horizontally using the scrollbar at the bottom of the screen to access the evaluation criteria.
- Filling Out the Sheet: To evaluate a task, click on the dropdown menu in the respective column under the evaluation criteria and select one of the available options. Column "S" is reserved for entering comments about the task.
- Instructions for Use:
 - Use the provided link to access the Google Sheets document.
 - Do not download the file; work directly in the online sheet optimised for the Google platform to prevent data loss.
 - Use a laptop or desktop computer rather than a mobile device for better functionality.
 - Ensure an internet connection to allow automatic saving of progress.

The second part of the instructions detailed the task evaluation criteria. The evaluation criteria were modeled after TIMSS item writing guidelines (Mullis et al., 2017). The exact wording of the criteria was as follows:

1. Content Domain: Assess the scope of knowledge and skills measured by the task, choosing one of the following categories: Numbers, Geometry and Measurement, Data, or None of the Above (for maths) or Life Science, Physical Science, Earth Science or None of the Above (for science).

2. Cognitive Domain (Skills): Determine the cognitive processes measured by the task, selecting one of the following options: Knowing, Applying, Reasoning or None of the Above.
3. Difficulty (Level Suitability): Evaluate whether the task is appropriate for the target group (grades IV-V), considering cognitive development rather than strict alignment with the national curriculum:
 - Yes: The task is of suitable difficulty.
 - No - Too Easy: The task is too easy.
 - No - Too Difficult: The task is too difficult.
4. Correctness: Assess the mathematical (scientific) correctness of the task:
 - Yes: The task is correct.
 - No: The task is incorrect.
5. Language: Evaluate whether the language used in the task is correct and natural:
 - Yes: The language is appropriate.
 - No: The language is inappropriate.
6. Clarity: Assess whether the task is clear and if the student will understand what is required:
 - Yes: The task is clear.
 - No: The task is unclear.
7. Attractiveness: Rate the task's attractiveness and engagement level for students:
Scale from 1 to 5, where 5 is very attractive and 1 is not attractive at all.
8. Usefulness: Evaluate the usefulness of the task in a school test:
 - Ready to Use: The task is suitable for use.
 - Needs Modification: The task needs modifications (e.g., changing an option, diagram, or wording).
 - Not Suitable: The task is not suitable for use.

For multiple-choice (MC) items evaluators were asked to provide some additional evaluation criteria:

9. One Correct Answer: Assess whether the task contains only one correct answer:
 - Yes: There is only one correct answer.

- No: There are multiple correct answers or none are correct.
10. Justified Answers: Evaluate whether all incorrect answer options are pedagogically justified (represent common errors or misconceptions):
- Yes: All options are justified.
 - No: Not all options are justified.
11. Answer Without Reading: Evaluate whether students can provide the correct answer to the task without reading the answer choices. Select one of the options:
- Yes
 - No

(This criterion was used for science items only. Its role is to control if the stem is clear and the whole item unambiguous.)

12. Comments: Provide comments if any aspect of the task catches your attention or if you want to share reflections on the task. Comments are not mandatory for every task but are encouraged when necessary.

3.2 Subject matter expert panel composition

Each set of items, the maths and the science items, were evaluated by five subject matter experts (SMEs). All SMEs were experienced mathematics or science teachers with on average 30 (maths) or 20 (science) years of teaching experience. All of them possessed at least a master's degree in maths or science (biology, geography or chemistry) and a significant additional experience in educational assessment. The most important details on the SMEs are presented in Table 1.

Table 1*Key information on subject matter experts*

Rater	Domain	Degree	Years of experience	Additional experience	Age	Gender
#1	maths	MA	35	Examiner for Polish high-stakes national assessments	60	female
#2	maths	MA	27	Examiner for Polish high-stakes national assessments	51	female
#3	maths	MA	37	Examiner for Polish high-stakes national assessments; Author of handbooks	61	female
#4	maths	MA	35	Examiner for Polish high-stakes national assessments	59	female
#5	maths	MA	16	Examiner for Polish high-stakes	40	female

				national assessments		
#6	science	PhD (biology)	16	Examiner for Polish high-stakes national assessments; Author of handbooks	48	female
#7	science	MA (geography)	29	Author of handbooks and low-stakes tests for educational publishers	55	female
#8	science	MA (biology)	8	Examiner for Polish high-stakes national assessments	53	female
#9	science	MA (biology)	24	Examiner for Polish high-stakes national assessments	66	female
#10	science	MA (biology/chemistry)	22	Author of handbooks and low-stakes tests for educational	47	female

				publishers		
--	--	--	--	------------	--	--

3.3 SMEs panel results

3.3.1 Agreement Analyses

To assess inter-rater agreement, we considered a range of widely used coefficients: Percent Agreement, Brennan and Prediger’s coefficient, Cohen’s/Conger’s Kappa, Scott’s/Fleiss’ Kappa, Krippendorff’s Alpha, and Gwet’s AC. Each of these metrics has distinct statistical properties, assumptions, and sensitivities, making them more or less suitable depending on factors such as the number of raters, number of categories, measurement scale, and the marginal distribution of ratings. Importantly, there is no consensus in the literature regarding which agreement coefficient should be used as a gold standard. A lively debate continues around the comparative merits of some of the coefficients, for instance - Gwet’s AC versus Cohen’s Kappa. Some studies point to the greater stability and interpretability of Gwet’s approach (e.g., Wongpakaran et al., 2013), while others argue in favor of the traditional Kappa framework (e.g., Vach & Gerke, 2023), particularly under conditions of well-balanced distributions.

In our analyses, we selected Gwet’s AC as the primary indicator of inter-rater agreement (Gwet, 2014). This coefficient was developed as a more stable alternative to Kappa-type statistics, particularly in scenarios where the marginal distributions are skewed. Gwet’s AC is less prone to what is known as the "Kappa paradox" — situations in which observed agreement is high but Kappa yields a low value due to statistical artifacts.

For additional context, we report Percent Agreement as a supplementary, descriptive measure. While it simply reflects the proportion of times raters assigned the same category, it does not account for chance agreement and is highly sensitive to the number of rating categories. In binary settings, Percent Agreement can appear

deceptively high due to chance alone, while in settings with more categories, it may underestimate true agreement.

The results of the raters' agreement are presented in Tables 2 and 3, for mathematics² and science items, respectively.

Table 2

Rater Agreement for Math Items

	<i>Percent Agreement</i>	<i>Gwet's AC</i>	<i>Rating categories</i>
<i>Correctness</i>	<i>0.88</i>	<i>0.85</i>	<i>2</i>
<i>Language</i>	<i>0.76</i>	<i>0.68</i>	<i>2</i>
<i>Clarity</i>	<i>0.82</i>	<i>0.77</i>	<i>2</i>
<i>Usefulness</i>	<i>0.55</i>	<i>0.38</i>	<i>3</i>
<i>Difficulty</i>	<i>0.74</i>	<i>0.68</i>	<i>3</i>
<i>Content congruence</i>	<i>0.97</i>	<i>0.97</i>	<i>2</i>
<i>Cognitive congruence</i>	<i>0.75</i>	<i>0.49</i>	<i>2</i>
<i>Justified options</i>	<i>0.78</i>	<i>0.70</i>	<i>2</i>
<i>One correct answer</i>	<i>0.92</i>	<i>0.90</i>	<i>2</i>
<i>Attractiveness</i>	<i>0.37</i>	<i>0.24</i>	<i>5</i>

Note. Number of raters = 5.

² One item, item M090067, was deemed “unfit to be rated” by one of the SMEs. Based on the teacher’s comment that the shape described in the exercise was not specified, the values for the “Clarity”, “Usability”, and “Merit based correctness” of the item were coded as “Not clear”, “Not usable”, and “Not correct” respectively. The rest of the values (“Cognitive Type”, “Difficulty”, “Language”, and “Attractiveness”) were impossible to determine from the comment of the teacher rater, and thus were not taken into account in the analysis.

Table 2 presents inter-rater agreement for ten rating criteria using Gwet's AC as the primary measure, along with Percent Agreement reported for reference. The number of rating categories varied across criteria, ranging from two (binary decisions) to five. The highest levels of agreement were observed for binary-coded items such as *Content congruence* (AC = 0.97), *One correct answer* (AC = 0.90), and *Correctness* (AC = 0.85). Other criteria such as *Clarity* (AC = 0.77), *Language* (AC = 0.68), and *Justified options* (AC = 0.70) also demonstrated substantial agreement. Agreement was lower for more complex or subjective criteria, particularly those with more response categories. For instance, *Usefulness* and *Difficulty* (both with 3 categories) yielded AC values of 0.38 and 0.68, respectively. While *Difficulty* still reached a substantial level of agreement, *Usefulness* fell to the lower range, indicating greater divergence in raters' judgments. *Cognitive congruence*, which had a high Percent Agreement of 0.75, but a much lower Gwet's AC of 0.49, pointing to only moderate true agreement after accounting for chance.

The lowest agreement was observed for *Attractiveness* (5 categories), where Gwet's AC dropped to 0.24 — indicating much lower agreement than in case of other criteria. This likely reflects the increased subjectivity and scale granularity, which tend to reduce consistency among raters. This low value may also reflect that the instruction did not explain all of the criteria clearly enough.

Table 3

Rater Agreement for Science Items

	<i>Percent Agreement</i>	<i>Gwet's AC</i>	<i>Rating categories</i>
<i>Correctness</i>	<i>0.81</i>	<i>0.75</i>	<i>2</i>
<i>Language</i>	<i>0.73</i>	<i>0.61</i>	<i>2</i>
<i>Clarity</i>	<i>0.73</i>	<i>0.59</i>	<i>2</i>
<i>Usefulness</i>	<i>0.45</i>	<i>0.20</i>	<i>3</i>

<i>Difficulty</i>	0.65	0.54	3
<i>Content congruence</i>	0.97	0.97	2
<i>Cognitive congruence</i>	0.51	0.03	2
<i>Justified options</i>	0.55	0.13	2
<i>One correct answer</i>	0.80	0.71	2
<i>Answer w/o reading</i>	0.72	0.51	2
<i>Attractiveness</i>	0.31	0.15	5

Note. Number of raters = 5.

The results for science items (Table 3) show that inter-rater agreement varies substantially across rating criteria, indicating differences in how clearly and consistently each property of the items was evaluated. Using Gwet's AC as the primary indicator of reliability, we observed high agreement for certain well-defined aspects, while more subjective or interpretive dimensions showed low consistency among raters. The highest agreement was found for *Content congruence* (AC = 0.97), suggesting that raters were nearly unanimous in determining whether an item was aligned with the science curriculum. Similarly, *Correctness* (AC = 0.75) and *One correct answer* (AC = 0.71) achieved substantial agreement, indicating that these task characteristics are well operationalized and reliably judged. Lower agreement was found for *Language* (AC = 0.61), *Clarity* (AC = 0.59), and *Difficulty* (AC = 0.54), suggesting that raters generally agreed on these aspects, although there is still room for increased standardization or clearer guidance in applying these criteria. In contrast, very low agreement was observed for *Usefulness* (AC = 0.20), *Justified options* (AC = 0.13), and especially *Cognitive congruence* (AC = 0.03). These values indicate a high degree of subjectivity or ambiguity in how these dimensions were interpreted and applied by different raters, casting doubt on their reliability without further clarification or rubric refinement. *Attractiveness*, which used a five-point scale, also showed poor agreement (AC = 0.15), likely due to increased

subjectivity introduced by the scale's granularity or by confusing instructions. Although Percent Agreement was sometimes higher (e.g., 0.72 for "Answer without reading"), the corrected measure (Gwet's AC) reveals that much of this agreement may be due to chance. Additionally, it could be noted that, it is easier to encounter more diverse opinions in a larger group of raters, which also lowers values of the agreement indicators but does not make low agreement more acceptable (Falotico & Quatto, 2015; Willerman, 1955).

3.3.2 Raters' evaluations

For the questions evaluated using binary judgments, we calculated the percentage of "positive" responses (e.g., selecting "The item's language is clear" under the Language criterion) and compared these ratings between the original TIMSS items and those generated by the LLM under different prompting conditions (zero-shot, one-shot, three-shot, and five-shot). To assess statistical significance, we applied a multilevel regression model with a cross-classified random effects structure, treating both raters and items as random factors. This approach allowed us to account for the hierarchical nature of the data, capturing variance attributable to both sources independently. We used the REML (Restricted Maximum Likelihood) estimator, which provides more accurate parameter estimates in cases with a limited number of groups (Raudenbush & Bryk, 2002). The criteria evaluated included: Correctness, Language, Clarity, Content Domain Congruence, Cognitive Domain Congruence, Justified Opinion, and Presence of One Correct Answer.

For items with multiple discrete answer options we have plotted the results on a categorical percentage heatmap. These criteria were: Usefulness, and Difficulty. For the Attractiveness, the only criterion rated on a truly ordinal answer scale, we have opted for a bar chart of means, with 95% confidence intervals, and a violin plot to inquire into the answer distributions.

3.3.2.1 Mathematics

Table 4 presents a detailed comparison of the rated quality of items across different setups, including the original TIMSS items and those generated by the LLM in

zero-shot, one-shot, three-shot, and five-shot conditions. The comparison focuses on three key criteria: Correctness, Language, and Clarity. For each criterion, the table provides the mean ratings, the differences relative to the TIMSS items, and the corresponding p -values for statistical significance.

Table 4

Ratings for Correctness, Language and Clarity - Mathematics

	Correctness			Language			Clarity		
	Mean	Dif.	p -val	Mean	Dif.	p -val	Mean	Dif.	p -val
TIMSS	.94			.84			.86		
Zero-shot	.85	-.09	.086	.83	-.015	.791	.81	-.05	.410
One-shot	.91	-.03	.567	.86	.02	.727	.82	-.04	.510
Three-shot	.93	-.01	.849	.88	.04	.485	.94	.08	.187
Five-shot	.95	.01	.849	.85	.01	.862	.94	.08	.187

Note. All criteria were rated on a dichotomous $\{0,1\}$ scale, hence maximum score equals 1.

The number of examples provided to the LLM appears to have influenced the rated quality of merit-based correctness of the items. The fiveshot setup achieved the highest rated correctness, marginally surpassing even the original TIMSS items (95% vs. 94%). The zero-shot (85%) and oneshot (91%) setups yielded the lowest correctness ratings, with the three-shot (93%) and fiveshot (95%) setups being closer to the original TIMSS items in quality. Notably, the three-shot, fiveshot, and TIMSS items all performed at a comparable level in terms of correctness, with differences being statistically non-significant.

Regarding the naturalness and correctness of the item language, all setups, including TIMSS items, were rated similarly (83-88%), suggesting that the language quality of generated items is largely comparable to that of the original items. Among these,

the three-shot setup showed a marginally higher rating (88%), although the differences were not statistically significant.

In terms of clarity, there was more variation among the setups. The highest ratings were observed for the fiveshot and three-shot setups (94%), while the original TIMSS items followed as the second-best-rated group (86%). The oneshot (82%) and zero-shot (81%) setups performed the worst, with differences in clarity ratings between the setups showing marginal statistical significance.

Figure 4 presents Usefulness Ratings by Setup. In terms of usefulness for the test creators, the results are again varied, with the original TIMSS items garnering the most “The item is ready to use” labels. However, the three-shot setup seems to be only marginally worse, followed closely by the zero-shot setup. The two other setups fiveshot, and oneshot turned out to be the worst in this regard.

Figure 4

Usefulness Ratings by Setup – Mathematics

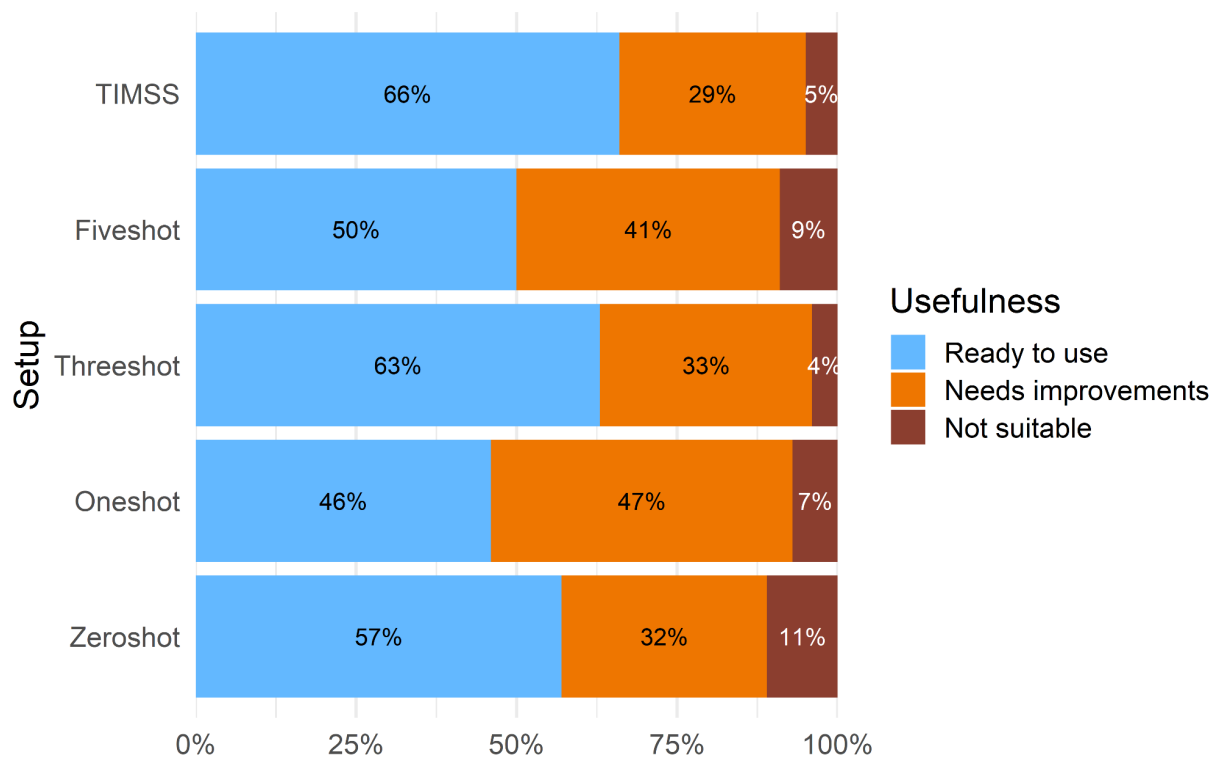
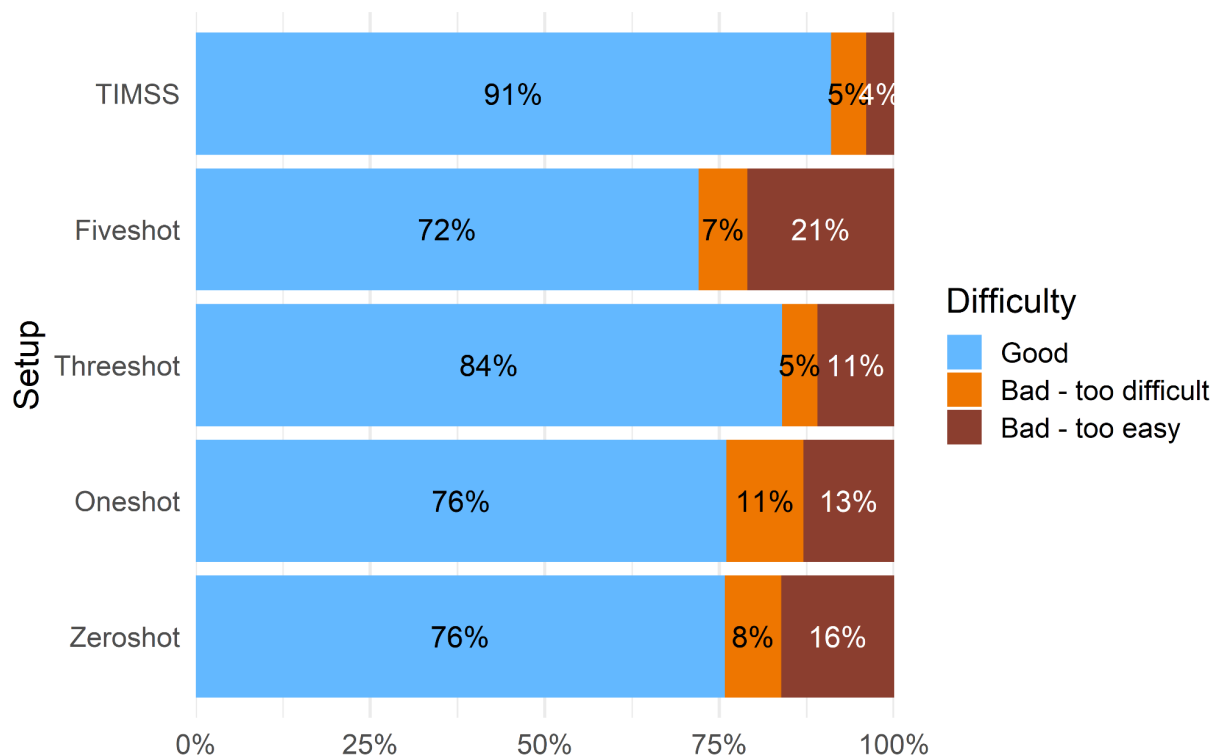


Figure 5

Difficulty Ratings by Setup – Mathematics



In terms of the Difficulty of the items, it seems again that the TIMSS items were evaluated as presenting the most suitable difficulty level, with the three-shot trailing closely after (similarly to the Usefulness criterion). Fiveshot, oneshot, and zero-shot turned out to be the worst in this regard. Interestingly, while for the TIMSS items the distribution of the two negative options was similar (5 % for “too difficult” and 4% for “too easy”), for the generated items teachers indicated more items to be “too easy”, than “too difficult” on average.

A simplified summary of the results from Figures 4 and 5 is presented in Table 5, where the indicators have been dichotomized. For Difficulty, a value of 1 was assigned to items deemed appropriately difficult, while others received a value of 0. Similarly, for Usefulness, tasks considered ready for use were assigned a value of 1, and others received a value of 0.

Table 5*Ratings for Usefulness and Difficulty - Mathematics*

	Usefulness			Difficulty		
	Mean	Dif.	<i>p</i> -val	Mean	Dif.	<i>p</i> -val
TIMSS	.66			.91		
Zero-shot	.57	-.09	.332	.76	-.15	.045
One-shot	.46	-.20	.031	.76	-.15	.047
Three-shot	.63	-.03	.746	.84	-.07	.354
Five-shot	.50	-.16	.085	.72	-.19	.012

Note. The table displays dichotomised usefulness and difficulty ratings, hence maximum score here is 1.

The original TIMSS items received the highest mean rating for usefulness (0.66), serving as the benchmark and reference category for further comparisons. Among the generated setups, the three-shot configuration performed closest to TIMSS, with a mean of 0.63, yielding only a negligible difference of -0.03 ($p = 0.746$). The five-shot setup followed with a mean of 0.50 (on the verge of statistical significance), and the zero-shot configuration scored slightly lower at 0.57. The one-shot setup had the lowest usefulness rating, with a mean of 0.46 and a statistically significant difference of -0.20 ($p = 0.031$).

For difficulty, the TIMSS items again achieved the highest mean rating (0.91). The three-shot setup was the next closest with a mean of 0.84, while the zero-shot and one-shot setups both scored similarly, but both exhibited statistically significant differences compared to TIMSS (-0.151, $p = 0.045$ for zero-shot, and -0.15, $p = 0.047$ for one-shot). The five-shot setup received the lowest difficulty rating, with a mean of 0.72 and a statistically significant difference of -0.19 ($p = 0.012$).

Table 6 presents comparisons of Content Congruence and Cognitive Congruence across different item generation settings, including the original TIMSS items and those generated by LLMs. This table is analogous to Table 4, providing a similar structure for evaluating how each setup aligns with the intended content and cognitive domains.

Table 6*Ratings for Content and Cognitive Congruence - Mathematics*

	Content Congruence			Cognitive Congruence		
	Mean	Dif.	<i>p</i> -val	Mean	Dif.	<i>p</i> -val
TIMSS	.88			.65		
Zero-shot	.99	.11	.028	.61	-.047	.697
One-shot	.98	.10	.045	.45	-.20	.096
Three-shot	1.0	.12	.016	.36	-.29	.016
Five-shot	.95	.07	.161	.54	-.11	.359

Note. The ratings code the mean congruence of teachers ratings of content and cognitive domains. SMEs had to choose among three options of different content and cognitive domains.

In terms of Content Congruence, the original TIMSS items received the lowest ratings among all setups, with a mean score of 0.88. The five-shot setup followed with a mean of 0.95, while the zero-shot, one-shot, and three-shot setups achieved near-perfect congruence ratings, with means of 0.99, 0.98, and 1.00, respectively. The differences between the TIMSS items and the zero-shot (0.11, $p = 0.028$), one-shot (0.10, $p = 0.045$), and three-shot (0.12, $p = 0.016$) setups were statistically significant, indicating higher congruence of raters ratings with regard to content domain in these generated items compared to the original TIMSS items.

Regarding Cognitive Congruence, the TIMSS items outperformed the generated items, with a mean score of 0.65. The three-shot setup received the lowest rating, with a mean of 0.36, followed by the one-shot (0.45), five-shot (0.54), and zero-shot (0.61) setups. The difference between the TIMSS items and the three-shot setup was statistically significant (-0.29, $p = 0.016$), suggesting that the original TIMSS items had higher congruence of raters ratings with regard to cognitive domain in these generated items compared to the original TIMSS items.

Table 7 presents the mean ratings for three criteria—Justified Options, One Correct Answer, and Attractiveness—comparing original TIMSS items with those generated by Large Language Models (LLMs) under zero-shot, one-shot, three-shot, and five-shot conditions.

Table 7*Ratings for Justified opinions, Presence of One Correct Answer and Attractiveness - Mathematics*

	Justified options			One correct answer			Attractiveness		
	Mean	Dif.	<i>p</i> -val	Mean	Dif.	<i>p</i> -val	Mean	Dif.	<i>p</i> -val
TIMSS	.873			.926			4.22		
Zero-shot	.872	-.002	.984	.905	-0.021	.908	3.68	-.54	0
One-shot	.744	-.126	.142	.899	-0.027	.911	3.72	-.50	.001
Three-shot	.908	.036	.681	.915	-0.011	.733	3.89	-.33	.034
Five-shot	.716	-.153	.077	.884	-0.042	.854	3.63	-.59	0

Note. Justified options and one correct answer were dichotomous variables - SMEs had to choose 0 or 1. Hence, maximum score equals 1 for these variables. Attractiveness was rated on a 5-point rating scale, hence maximum score equals 5.

In evaluating whether the available answer options (key and distractors) were justified, the three-shot setup achieved a mean rating of approximately 0.908, surpassing the original TIMSS items, which had a mean of 0.873. The zero-shot setup performed similarly to the TIMSS items, with a mean of 0.872. In contrast, the five-shot and one-shot setups received lower ratings, with means of 0.716 and 0.744, respectively, indicating a decline of about 10% compared to the TIMSS items - a difference on the verge of statistical significance. It means that, on average, SMEs rated response options as justified in 87% of cases among TIMSS items and 91% of three-shot LLM-generated items, but only in case of, on average, 72% in five-shot LLM-generated items.

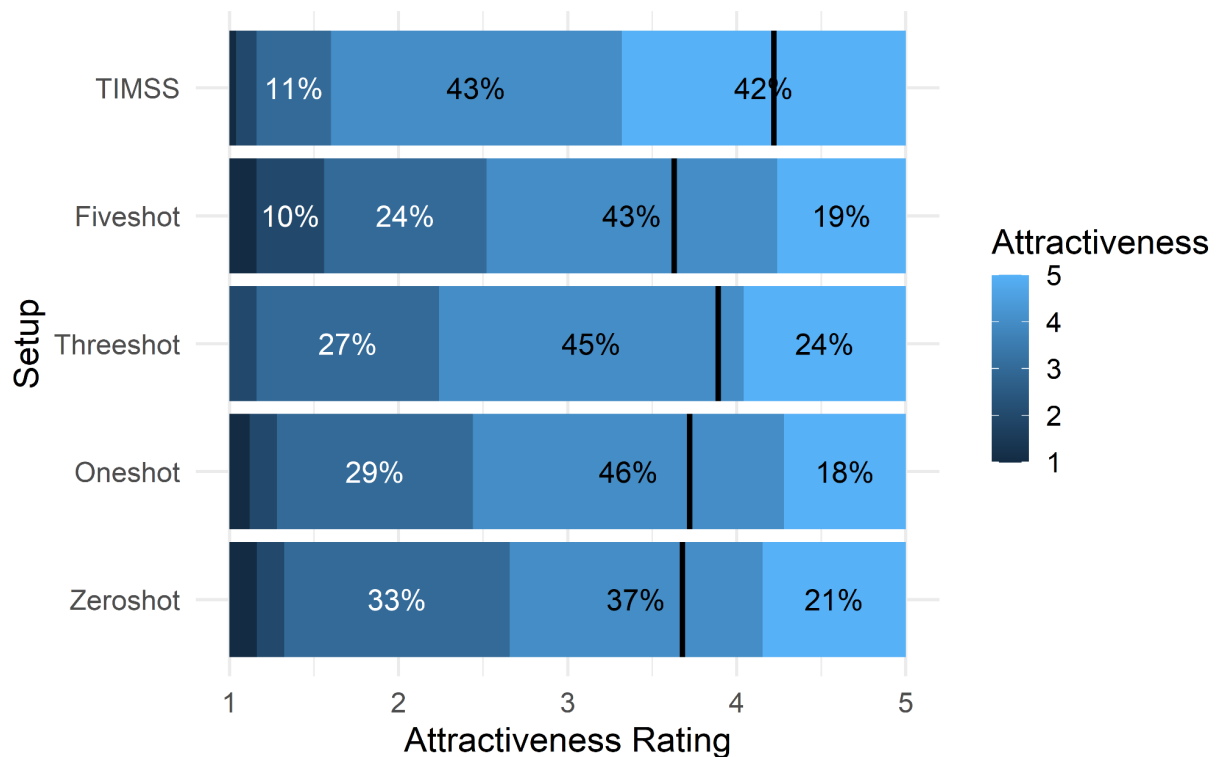
Regarding the criterion of having only one correct answer, all setups, including the TIMSS items, exhibited comparable performance, with mean ratings around 0.90, suggesting no significant differences across the different item generation methods.

In terms of Attractiveness, the TIMSS items received the highest mean rating of 4.22. All generated item setups scored lower and these differences were statistically

significant. The three-shot setup achieved a mean of 3.89, followed by the one-shot (3.72), zero-shot (3.68), and five-shot (3.63) setups. These differences indicate that the TIMSS items were perceived as more attractive compared to the generated items. The Attractiveness ratings are also illustrated in Figure 6, providing a visual comparison across the different setups. Figure 6 seems to suggest that TIMSS items especially distanced the LLM-generated items in case of maximum, 5-point ratings of Attractiveness.

Figure 6

Attractiveness Ratings Across Different Item Generation Setups



Note. Vertical bars indicate mean ratings.

In our analysis, the three-shot setup consistently outperformed other generated item configurations and, in certain aspects, even surpassed the original TIMSS items. Notably, in the Justified Options criterion, the three-shot setup achieved a mean rating of 0.908, slightly higher than the TIMSS mean of 0.873. Similarly, for Clarity, the three-shot setup received a mean score of 0.94, compared to the TIMSS mean of 0.86. However, an exception was observed in the Cognitive Domain Congruence, where the three-shot

setup had a mean rating of 0.36, significantly lower than the TIMSS mean of 0.65 ($p = 0.016$). Given that the TIMSS items themselves scored only 65% in this criterion, we consider this an outlier and recommend further investigation into the cognitive alignment of generated items.

These findings suggest that while the three-shot setup shows promise in enhancing certain item qualities, attention must be paid to cognitive domain alignment to ensure the effectiveness of generated assessment items.

3.3.2.2 Science

We analyzed the science items in a manner analogous to the mathematics items. Table 8 presents the results for Correctness, Language, and Clarity for the science items.

Table 8

Ratings for Correctness, Language and Clarity - Science

	Correctness			Language			Clarity		
	Mean	Dif.	p -val	Mean	Dif.	p -val	Mean	Dif.	p -val
TIMSS	.92			.80			.82		
Zero-shot	.84	-.08	.195	.81	.01	.881	.81	-.01	.899
One-shot	.82	-.10	.105	.78	-.02	.765	.73	-.09	.252
Three-shot	.87	-.05	.418	.85	.05	.454	.77	-.05	.524
Five-shot	.895	-.025	.685	.829	.029	.665	.82	-.02	.796

Note. All criteria were rated on a dichotomous (0-1) scale, hence maximum score equals 1.

In terms of merit-based correctness, the TIMSS items achieved the highest score (0.92). Among the generated setups, the five-shot configuration performed the best (0.895), followed by the three-shot (0.87), zero-shot (0.84), and one-shot (0.82) setups. While the differences between these setups were not statistically significant ($p > 0.1$),

they reflect a gradual decline in correctness from the TIMSS items to the one-shot configuration. However, as all setups scored between 80% and 90%, the differences can be treated as marginal.

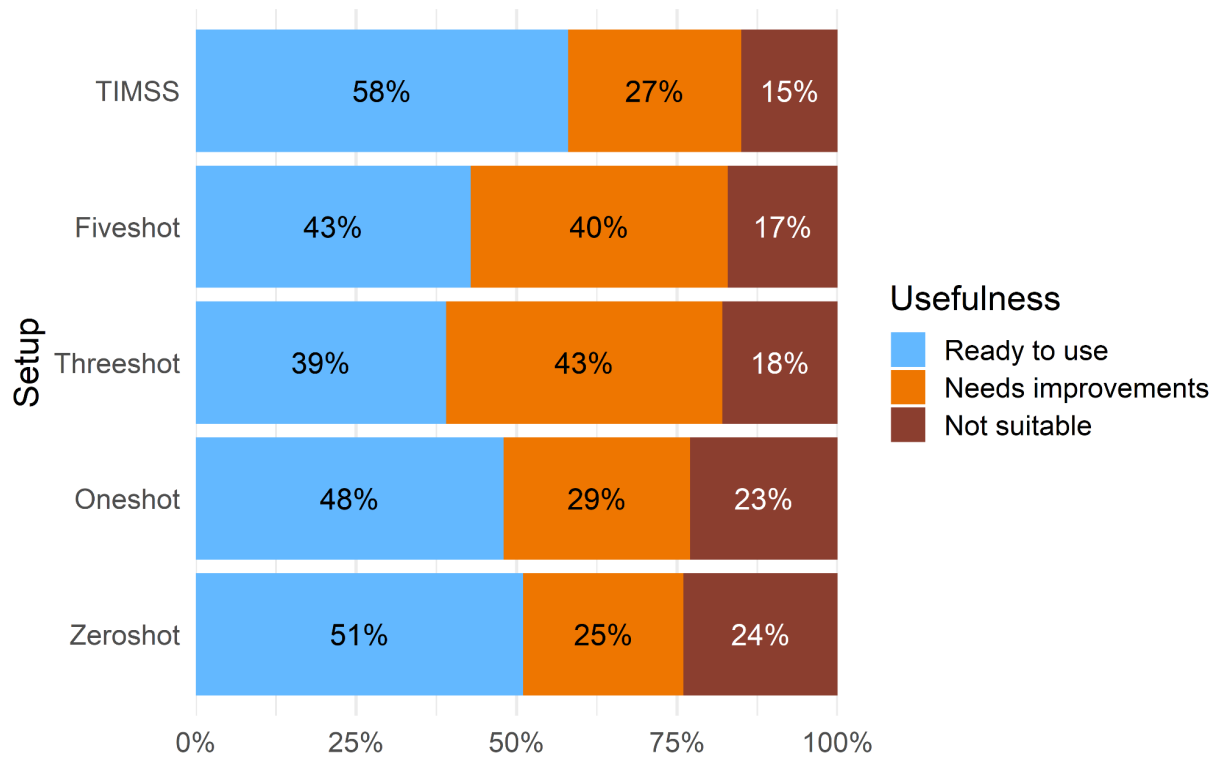
Regarding language clarity and naturalness, both the three-shot (0.85) and five-shot (0.83) setups marginally outperformed the original TIMSS items (0.80). The zero-shot setup scored similarly to the TIMSS items (mean = 0.81, $p = 0.881$), while the one-shot setup had the lowest score (0.78), although this difference was not statistically significant ($p = 0.765$). These findings suggest that the generated items generally maintain comparable or slightly better language clarity compared to the original items.

For clarity, the TIMSS and five-shot setups tied for the highest mean score (0.82), with the zero-shot setup scoring similarly (mean = 0.81, $p = 0.899$). The three-shot setup scored slightly lower (0.77), and the one-shot setup had the lowest clarity score (0.73). While none of the differences reached statistical significance ($p > 0.1$), the results highlight a consistent trend where the TIMSS items match five-shot and zero-shot and only very slightly outperform three-shot in terms of clarity, with the exception of the one-shot configuration, which lags behind more noticeably.

Figure 7 presents the Usefulness Ratings by Setup, while Figure 8 illustrates the Difficulty Ratings by Setup. For the purpose of regression analysis, usefulness and difficulty ratings were dichotomised, as shown in Table 9, where items labeled as “Ready to Use” and with appropriate difficulty were assigned a value of 1, and others were assigned a value of 0.

Figure 7

Usefulness Ratings by Setup – Science Items



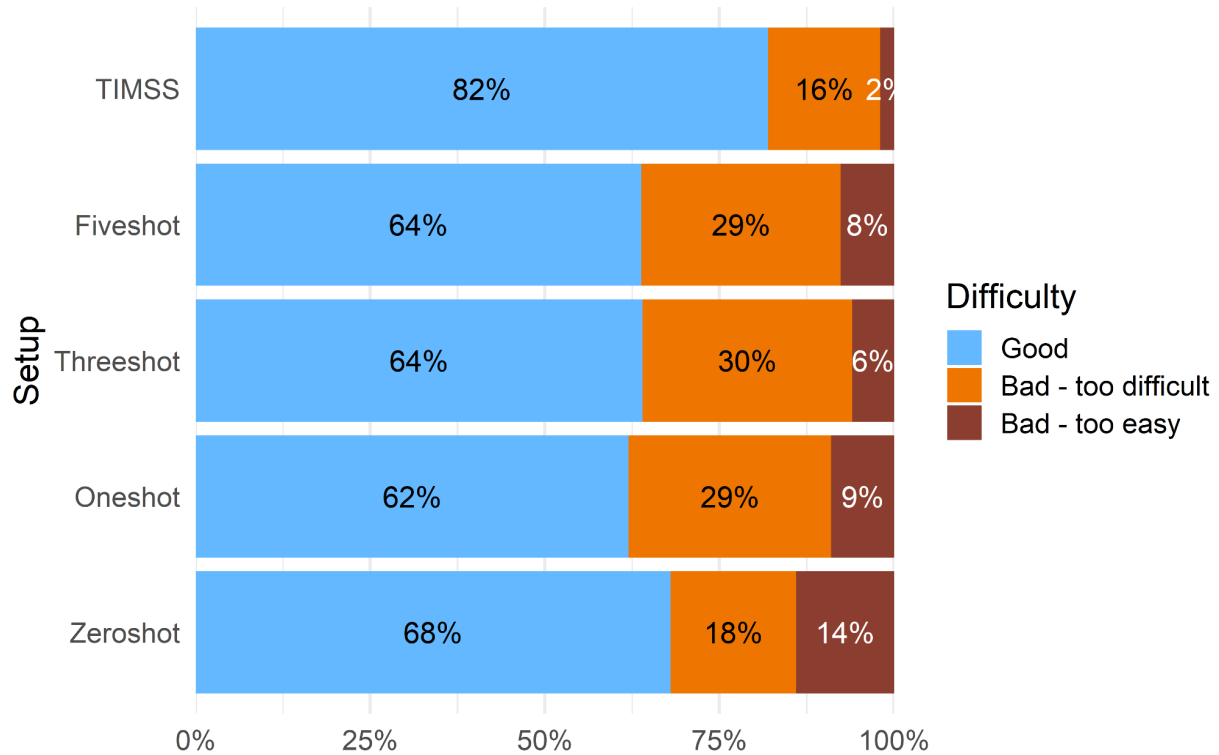
Note. The graph presents the percent (%) of the items in a given condition that were classified to a given usefulness category (“Ready to use”, “Needs improvement”, “Not suitable”).

In terms of usefulness for test use, the TIMSS items achieved the highest mean rating (0.58), serving as the benchmark. Among the generated setups, the zero-shot configuration performed the best (mean = 0.51, $p = 0.451$), followed by the one-shot setup (mean = 0.48, $p = 0.282$) and the five-shot (mean = 0.429, $p = 0.099$). These setups showed a marginal decrease compared to the TIMSS items but were not significantly worse.

However, the three-shot setup (mean = 0.39, $p = 0.041$) performed statistically significantly worse than the TIMSS items. These findings highlight a notable drop in perceived readiness for use as the number of examples increases beyond one-shot generation. However, it should be noted that three- and five-shot setups noted fewer items rated as “not suitable” than zero- and one-shot setups.

Figure 8

Difficulty Ratings by Setup – Science Items



Note. The graph presents the percent (%) of the items in a given condition that were classified to a given difficulty category (“Good”, “Bad - too difficult”, “Bad - too easy”).

For difficulty, the TIMSS items again outperformed all other setups, with 82% of items deemed as of good level of difficulty for a grade 4 test. Among the generated setups, the zero-shot configuration performed the best (mean = 0.68, $p = 0.122$) but still fell short of the TIMSS benchmark. The one-shot (mean = 0.62, $p = 0.027$), three-shot (mean = 0.64, $p = 0.047$), and five-shot (mean = 0.638, $p = 0.042$) setups performed similarly to the zero-shot, but all significantly worse than the TIMSS items.

Interestingly, all setups, including TIMSS, received more "Too Difficult" labels than "Too Easy" labels. This suggests a consistent trend where items are perceived as challenging, with a relatively low proportion of items being deemed too simple. Zero-shot setup generated by far the largest proportion of such items.

Table 9*Ratings for Usefulness and Difficulty - Science*

	Usefulness			Difficulty		
	Mean	Dif.	p-val	Mean	Dif.	p-val
TIMSS	.58			.82		
Zero-shot	.51	-.07	.451	.68	-.14	.122
One-shot	.48	-.1	.282	.62	-.2	.027
Three-shot	.39	-.19	.041	.64	-.18	.047
Five-shot	.429	-.151	.099	.638	-.182	.042

Note. The table displays dichotomised usefulness and difficulty ratings, hence maximum score here is 1.

The next table, Table 10, presents the results for Content Domain Congruence and Cognitive Domain Congruence. The content domain congruence did not differ significantly between the setups, with all configurations scoring near 100% on average. The TIMSS items scored a mean of 0.98, which was closely matched by the three-shot setup (mean = 0.98, $p = 1$) and the zero-shot setup (mean = 1, $p = 0.544$). The one-shot (0.96) and five-shot (0.962) setups also performed similarly, with differences that were not statistically significant ($p > 0.5$).

Table 10.*Ratings for Content and Cognitive Congruence*

	Content congruence			Cognitive congruence		
	Mean	Dif.	p-val	Mean	Dif.	p-val
TIMSS	.98			.39		
Zero-shot	1.0	.02	.544	.52	.13	.076
One-shot	.96	-.02	.544	.49	.10	.173
Three-shot	.98	0.0	1	.42	.03	.683
Five-shot	.96	-.018	.578	.47	.08	.290

Note. The ratings code the mean congruence of teachers ratings of content and cognitive domains. SMEs had to choose among three options of different content and cognitive domains.

For cognitive domain congruence, there were more notable differences. Interestingly, despite the generation prompt providing math-based definitions of the cognitive domain, the original TIMSS items scored the lowest among all setups, with a mean of 0.39. The three-shot setup performed slightly better (mean = 0.42, $p = 0.683$), but still scored relatively low.

The one-shot (mean = 0.49, $p = 0.173$) and five-shot (mean = 0.47, $p = 0.29$) setups performed comparably, showing some improvement over the TIMSS items but not significantly. The best-performing configuration was the zero-shot setup, which achieved a mean score of 0.52, reflecting a 13% improvement over the TIMSS items ($p = 0.076$), though this difference was on the verge of statistical significance.

These results mean that raters were very congruent when assessing items' content domain, but they were much less congruent when assessing their cognitive domain. It seems that raters were less congruent when assessing science items in comparison to mathematics items in this criterion (see Table 6). This might be due to the mistake in the prompting scheme, where mathematics cognitive process definitions were mistakenly used to generate science items. Another possibility is that science items are inherently more difficult to classify into cognitive domains or that science cognitive domains are more difficult to classify. Yet another possibility is that science is a heterogeneous domain in TIMSS and poses more problems for congruent rating for raters of equally heterogeneous backgrounds (chemistry, biology, geography, etc.), making this group significantly different from the much more unified group of mathematics raters.

Table 11 contains information on Justified Options, One Correct Answer, and Answer Without Reading ratings.

Table 11

Ratings for Justified Options, Presence of One Correct Answer, and Answer without Reading - Science

Justified options	One correct answer	Answer w/o reading
-------------------	--------------------	--------------------

	Mean	Dif.	<i>p</i> -val	Mean	Dif.	<i>p</i> -val	Mean	Dif.	<i>p</i> -val
TIMSS	.706			.800			.382		
Zero-shot	.625	-.044	.515	.844	.084	.324	.156	-.246	.009
One-shot	.667	-.003	.966	.906	.146	.088	.240	-.175	.063
Three-shot	.674	0	.995	.840	.086	.317	.304	-.103	.279
Five-shot	.670	-.002	.974	.830	.055	.515	.255	-.139	.139

Note. Justified options, one correct answer and possibility to answer without reading the response options were all dichotomous variables - SMEs had to choose 0 or 1. Hence, maximum score equals 1 for these variables.

In terms of whether the response options for MC items were justified, the TIMSS items achieved the highest score (0.706), serving as the benchmark. The zero-shot setup scored the lowest (0.625), followed by the five-shot and one-shot setups. The three-shot configuration performed marginally better, but overall, the differences between the setups were not statistically significant ($p > 0.1$).

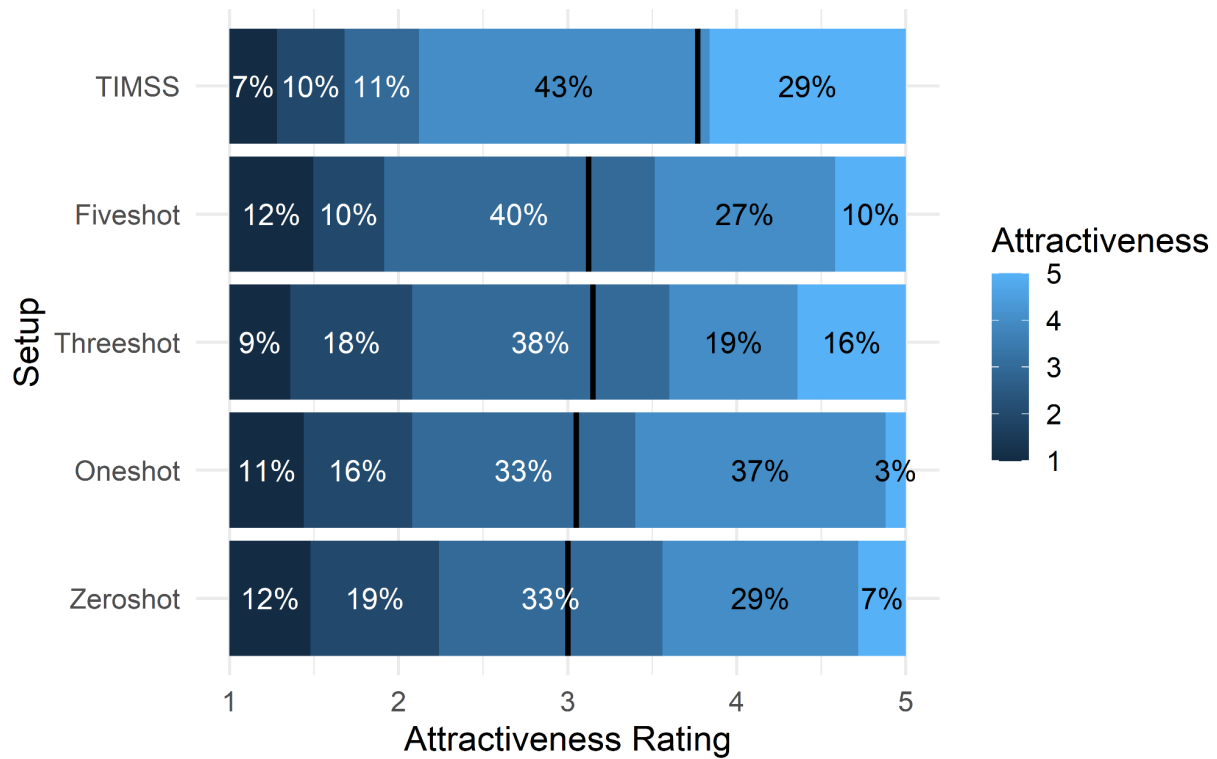
For one correct answer criterion, the TIMSS items scored the lowest among all setups (0.80). All generated setups, except for the one-shot, performed marginally better, although these differences were not statistically significant. Interestingly, the one-shot setup achieved the highest score (mean = 0.906, difference = 0.146, $p = 0.088$), indicating an improvement over the TIMSS items on the verge of statistical significance.

For items possible to be answered correctly without reading the response options, which was reverse-coded (i.e., higher percentages indicate worse performance), the TIMSS items performed the worst, with a mean of 0.382. This indicates that TIMSS items required the least effort for test-takers to determine the correct answer without engaging with the response options. Among the generated setups, the zero-shot configuration noted a significantly lower mean of 0.156 (difference = -0.246, $p = 0.009$). The one-shot setup scored 0.24 ($p = 0.063$), while the three-shot and five-shot setups scored 0.304 ($p = 0.279$) and 0.255 ($p = 0.139$), respectively, placing them in the middle of the range.

Finally, Figure 9 and Table 12 present the Attractiveness Ratings Across Different Item Generation Setups.

Figure 9

Attractiveness Ratings Across Different Item Generation Setups



Note. Vertical bars indicate mean ratings.

In terms of overall attractiveness, the original TIMSS science items were rated significantly higher than all other setups, with a mean attractiveness score of 3.77. This score highlights that TIMSS items are perceived as more engaging and appealing compared to those generated by Large Language Models (LLMs). The latter gathered especially few 4- and 5-point ratings.

Among the generated setups, the zero-shot configuration received the lowest rating (mean = 3.00, difference = -0.77, $p < 0.001$), closely followed by the one-shot setup (mean = 3.05, difference = -0.72, $p < 0.001$). The three-shot setup (mean = 3.15, difference = -0.62, $p = 0.002$) and five-shot setup (mean = 3.124, difference = -0.646, $p = 0.001$) are rated slightly better but still fell significantly short of the TIMSS items.

Table 12*Ratings for Attractiveness - Science*

	Mean	Dif.	Attractiveness <i>p</i> -val
TIMSS	3.77		
Zero-shot	3.00	-.77	0
One-shot	3.05	-.72	0
Three-shot	3.15	-.62	.002
Five-shot	3.124	-.646	.001

Note. Attractiveness was rated on a 5-point rating scale, hence maximum score equals 5.

3.3.2.3 Additional Analyses

To analyse whether the differences in Attractiveness and Usefulness could have been related to the quality of the images or the tables generated by the research team members with the use of LLMs, the dataset was filtered and these two analyses were recreated for the item sets that did not contain images or tables, which for science constituted 82 items, and for maths 69 items. The elimination of the items that included images or tables from the analysis did not affect the difference between the TIMSS and LLM items in a significant way.

We have then recreated the analysis for only those items that had images or tables. For science there were 19 such items and 31 for maths. Due to the very low number of items that included images or tables, we deem this specific analysis to be rather inconclusive, which is additionally evidenced in the width of the confidence intervals for Attractiveness ratings. No definite conclusions about the specificity of these items with regards to the rest of the item pool can be drawn.

3.3.3 Raters' comments

Raters' comments are important sources of qualitative data about item quality and reasons for their misfit for grade 4 assessment.

Mathematics items

In the case of mathematics, raters noticed the following problems with the LLM-generated items:

- Occasionally items refer to terms or problems that are too difficult for grade 4 students, e.g. in the item presented below, grade 4 students are not deemed able to calculate the area of the circle:
 - *Which of the following figures has the largest surface area?*
 - *A) A rectangle with sides 4 cm and 5 cm*
 - *B) A square with sides 6 cm*
 - *C) A right triangle with legs 3 cm and 4 cm*
 - *D) A circle with a radius of 3 cm*
- Too easy, overly simplistic items, suitable for grade 2 or early grade 3, e.g.:
 - *Jack went to the store and bought a box of blocks. The box contains 3 rows of 4 blocks in each row.*
 - *How many blocks are in the box?*
 - *A) 7*
 - *B) 12*
 - *C) 16*
 - *D) 24*
- Problems with distractors - implausible response options or lack of justified response options, based on common mistakes, e.g., referring to the task below, one of the experts wrote: *Instead of 52 and 62 I would write e.g. 43 and 41, I think that the possible mistakes of the students would be more predictable in such a case*
 - *What is $15 + 27$?*
 - *A) 32*
 - *B) 42*
 - *C) 52*
 - *D) 62*

- Another example: *Among the options I am missing the result 23, in my opinion students would most often make a mistake of not using the bracket, i.e. $7 \cdot 3 = 21$ and then $21 + 2 = 23$. Item:*
 - *Complete the following equation: $7 \cdot (3 + 2) = \underline{\hspace{2cm}}$*
 - *A) 21*
 - *B) 25*
 - *C) 35*
 - *D) 50*
- Problems with distractors - the same option (most often the correct option) is presented twice, e.g.:
 - *What is the sum of 123 and 456?*
 - *A) 579*
 - *B) 589*
 - *C) 579*
 - *D) 599*
- Problems with distractors - no correct response option, e.g., in the below item the correct response of 8 zlotys (PLN) seems to be missing:
 - *"Mrs. Kowalska went shopping and bought 3 kilograms of apples for 12 zlotys and 2 kilograms of pears for 8 zlotys. How much did she pay for the kilogram of apples and the kilogram of pears in total?"*
 - *A) 4 zlotys*
 - *B) 5 zlotys*
 - *C) 6 zlotys*
 - *D) 10 zlotys*
- Imprecise use of mathematical terms, e.g. use of "side" instead of "edge", in the context of solid figures.
- Imprecise use of other mathematical symbols, e.g. superfluous brackets:
 - *What number is the result of the expression $(32) - (2 \cdot 3)$?*
 - *A) 3*
 - *B) 9*
 - *C) 6*

- D) 0

Science items

In the case of science items, raters pointed to main such item flaws as:

- Improper or imprecise item formulation that hinders or excludes its proper solving, e.g.
 - *Which of the following phenomena is an example of a process occurring on the Earth's surface?*
 - *A) The formation of a hurricane*
 - *B) The temperature of the atmosphere rising*
 - *C) The melting of glaciers*
 - *D) Observing the stars in the night sky*
- Use of too difficult vocabulary or scientific terms that are not understandable by grade 4 students (e.g. reproduction, ecosystem, convection, kinetic energy, transpiration, molecules, etc.), e.g.
 - *The task requires reformulating vocabulary, students may not be familiar with the concept of ecosystem (grade 8 requirements), you can replace this word with environment or provide [an example of] a specific ecosystem, such as meadow, forest, lake;*
 - *At the level of grades 4-5, the concept of a molecule (not even a particle) is not used - we talk about substances in general.*
- Too easy items that are also overly simplistic and do not tap into any of the cognitive domains at grade 4 level, e.g.
 - *Which of the following animals is an example of a predator?*
 - *A) Rabbit*
 - *B) Cow*
 - *C) Wolf*
 - *D) Snail*
 - *The task, like many in this set, is simple, testing knowledge and does not contribute much to the learning process.*
- Problems with distractors, e.g. implausible response options, correct response option is obvious and very salient, e.g.

- *Which of the following is characteristic of living organisms?*
- *A) Respiration*
- *B) Reflection of light*
- *C) Formation of rocks*
- *D) Conduction of electricity*
- Further problems with response options, e.g. two or more answers are correct, or two response options are repeated, example of items with all response options correct:
 - *Which of the following processes is essential for plant survival?*
 - *A) Photosynthesis.*
 - *B) Cellular respiration.*
 - *C) Transpiration.*
 - *D) Pollination.*
 - *Which of the following devices uses electricity to operate?*
 - *A) A bicycle.*
 - *B) A refrigerator.*
 - *C) A car.*
 - *D) A gas stove.*
- Use of imprecise or improper terms, e.g.
 - *Water lifecycle*, instead of *water circulation*.
- Linguistic problems, e.g. occasional spelling, stylistic or grammar errors, superfluous words and other effects leading to linguistic awkwardness.

3.4 SMEs Panel Summary and Discussion

The analysis of mathematics items indicates that the three-shot setup emerged as the best-performing configuration among the LLM-generated items when compared to the original TIMSS items. In terms of clarity, the three-shot setup achieved the highest score (0.94), slightly outperforming the TIMSS items (0.86), although this difference was not statistically significant ($p = 0.187$). Similarly, for correctness, the three-shot setup performed on par with the TIMSS items (mean = 0.93 vs. 0.94, $p = 0.849$). Furthermore,

in the criterion of justified options, the three-shot setup demonstrated a slight advantage over TIMSS (mean = 0.908 vs. 0.873, $p = 0.681$), suggesting that SMEs found its distractors and correct options to be marginally better.

Despite these strengths, the TIMSS items maintained superiority in most areas. They were not only rated significantly higher in attractiveness (4.22) compared to the three-shot setup (mean = 3.89, $p = 0.034$), but also scored better in usefulness, being more frequently deemed "ready to use" for assessments (0.66 vs. three-shot 0.63, $p = 0.746$), and in difficulty, where they were rated as most appropriate for grade 4 test-takers (0.91 vs. three-shot 0.84, $p = 0.354$).

The analysis of science items does not reveal a clear winner among the different LLM-generated configurations. Depending on the criterion, different setups tend to perform better, but the differences between them are generally neither statistically significant nor substantively large. However, it is evident that LLM-generated items, while close in performance, are consistently rated slightly lower than the original TIMSS items across most criteria. Although these differences are small and only occasionally reach statistical significance, they highlight a systematic preference for the TIMSS items over those generated by LLMs in the context of science task

None of the generation conditions proved unequivocally or consistently superior. Items from certain conditions were rated higher on one criterion but lower on another. In both mathematics and science, raters noted problems with clarity and correctness, particularly for items generated under zero-shot and one-shot conditions.

In contrast, original TIMSS items were rated most favorably in terms of attractiveness, appropriate difficulty, and usefulness for real assessments. Interestingly, five-shot items tended to be rated slightly lower on these criteria for mathematics, though this pattern was less pronounced in science. For science items, one-shot items appeared to receive lower ratings, whereas in mathematics, both zero-shot and one-shot items tended to score lower.

An intriguing pattern emerged: LLM-generated mathematics items were generally seen as too easy, whereas science items were often considered too difficult. Additionally, the zero-shot condition produced the highest proportion of items deemed not suitable for

real assessments. Overall, science items received less favorable evaluations than mathematics items.

3.4.1 Limitations

A relatively low number of raters and items within condition is a limitation of this analysis, as more raters and a larger item pool could reveal additional insights. Another important limitation is that raters were working with the material using only the prepared instructional materials and asking occasionally for clarifications from the PI. Given the frames of the project there was no time for a proper raters calibration session, which may have affected the results, mainly with regard to inter-rater agreement. Low interrater-agreement undermines the validity of some of the indicators (e.g. Attractiveness for science items).

It is also probable that the online mode of the SMEs panel contributed to a lower agreement, as raters were unable to consult their opinions with each other. Maybe, the item rating sessions should be done once more, this time in a more classical way, with a stationary mode and with a proper rater calibration and training before rating session. This would allow us to compare the results between the two studies and identify true reasons for low inter-rater agreement in some of the criteria.

It is also noteworthy that none of our raters was ever involved in TIMSS studies, which may explain the low inter-rater congruence for assessing items' cognitive domains. TIMSS cognitive domains framework is specific for this assessment and it is not a widely accepted standard elsewhere. On the other hand, using raters not familiar with TIMSS assessment enabled to gather their unbiased opinions about the TIMSS items. However, using raters familiar with TIMSS, e.g. from The Science and Mathematics Item Review Committee (SMIRC), is an interesting idea for a future study.

3.4.2 Future directions

Future projects should continue with the SMEs evaluation of generated items, as their ratings and opinions provide a valuable source of information on items' quality and directions for further improvement. Few lines of such research are feasible. Small groups of SMEs with significant expertise in TIMSS can provide unique, more qualitative insight

on generated items quality (and usefulness for TIMSS assessment itself), but large groups of experts (> 50) can provide equally insightful, quantitative ratings about the items' most basic features (cf. Bezirhan & von Davier, 2023). As shown in this report, SMEs were able to identify most important problems with LLM-generated items, such as unattuned difficulty or low attractiveness. It is also important to employ verified samples of SMEs, as e.g. MTurk raters or intercept internet samples cannot be trusted with the level of their true qualifications and survey attentiveness (e.g. Kay, 2024).

3.4.3 Conclusions

Attempting a general summary, it appears that items generated by LLMs are slightly weaker than the original TIMSS items in most criteria. The TIMSS items generally outperform the LLM-generated items in terms of quality across the majority of criteria (or they are not worse), though not always in a statistically significant way. In some cases, the LLM-generated items even perform better in specific criteria, while it is clear that the TIMSS items excelled in some criteria, such as attractiveness, difficulty, and usefulness for assessment.

Finally, it is important to remember that this study compares the best-case items (original TIMSS) with the worst-case scenario items (automatically generated), which were not human-edited before entering the SMEs panel. Considering that TIMSS test items underwent a lengthy process of construction, piloting, and validation before their use in the main study, the fact that their advantage over the LLM-generated items is relatively minor suggests that LLMs represent a promising direction for item generation. Taking this into account, points to a rather high quality of the automatically generated items in this study, as rated by the SMEs.

4. Validation study

Empirical comparison of LLM-generated and original TIMSS items constituted a second step of LLM-generated item quality. In this part of the study we have generated a new set of items that were then mixed with original TIMSS items in test booklets. We then organised a fieldwork study in which more than 1200 students from Grade 4 took part.

The analysis of this study results concentrates on psychometric quality of the LLM-generated items and any differences between the generated and original TIMSS items.

4.1 Item generation

4.1.1 Overall Approach

Leveraging the results of the SMEs panel we considered two new item generation setups for the empirical study. One, informed by the pilot study – referred to from now on as the *simple* setup, and a second, prompted to specifically create either multiple choice or open-ended items, as well as to create items of slightly higher difficulty, referred to from now on as the *complex* setup. Both setups utilized the “gpt-4o” OpenAI API endpoint for generation. The two setups were used to generate 80 items each, distributed equally between the two domains of science and mathematics. The hyperparameters were kept the same as for the previous generation process: temperature = 0.8, top_p = 1, n = 1, presence_penalty = 0, frequency_penalty = 0, and max_tokens was left unspecified, allowing the model to generate up to the maximum possible length permitted by the context window. On the basis of the SMEs panel results, we have employed a three-shot prompting scheme to generate the items here.

4.1.2 The Simple Setup

The first setup, modelled after the setups from the pilot study, used the following query:

```
“Create one item for the Polish TIMSS assessment. Do not write anything else than the content
of the item. The item has to be in Polish.
Subject : {subject}
Content Domain : {content_domain}
Definition: {content_domain_definitions[content_domain]}
Cognitive Domain : {cognitive_domain}
Definition: {cognitive_domain_definitions[cognitive_domain] + f'_{subject[0]}'}
Age appropriateness: Grade 4 of the elementary school
Item format: Constructed response OR multiple-choice. Constructed response items usually
```

require students to give a numerical result, provide a short explanation or description given in one or two phrases or sentences, create a list or complete a table. Multiple choice items have four response options, ONE OF THEM IS ALWAYS CORRECT AND THREE ARE INCORRECTt. The incorrect options need to be plausible choices, demonstrating typical errors or misunderstandings. They need to be similar to the correct option in terms of length and format, but be undoubtedly incorrect.

Additional comments: DO NOT GENERATE LINKS TO TABLES AND IMAGES, BUT PRESENT A DETAILED DESCRIPTION OF TABLES AND IMAGES IF YOU CHOOSE TO USE THEM IN YOUR ITEM (YOU DONT HAVE TO)"

The final version of the query was arrived at through an iterative process of trial and error. As can be seen, some sentences are fully capitalized, reflecting the failures of addressing the LLM in normal tone. Since the capitalization achieved its intended aim, it was retained.

Where the Subject took the form of either “Mathematics”, or “Science”. The rest of the placeholder values were defined in the next section. Being able to specify the subject, cognitive, and content domain allows for guided generation of items from specific domains and subjects. This utility was used to generate a domain item distribution that complied with the distribution of original TIMSS items with regard to these characteristics.

To achieve this, first the content, and cognitive domain were sampled from the list of real cognitive and content domains tuples extracted from 352 TIMSS items coming from the 2011, 2015, and 2019 TIMSS editions. Next random items that shared the sampled cognitive and content were picked from among those scraped from TIMSS booklets. These items, along with the descriptions of domains outlined in the next section, were then embedded with the above query to form the generation query. Three items were presented as examples each time – a choice informed by the results of the SMEs panel described above.

The exemplary items were added to the end of the query after the snippet below:

"Examples of items that meet the requirements: "

Using the following format:

####

Example {example_number}:

{item text}

The query was then followed by:

####

Generated item: “

This was done to signal to the LLM that it should generate a novel item.

4.1.2.1 Cognitive Domains

The following definitions were used for TIMSS Grade 4 cognitive domains:

Knowing (Math):

Facility in applying mathematics, or reasoning about mathematical situations, depends on familiarity with mathematical concepts and fluency in mathematical skills. The more relevant knowledge a student is able to recall and the wider the range of concepts he or she understands, the greater the potential for engaging with a wide range of problem situations. It contains skills like: Recall: Recall definitions, terminology, number properties, units of measurement, geometric properties, and notation (e.g., $a \times b = ab$, $a + a + a = 3a$). Identify: Identify numbers, expressions, quantities, and shapes. Recognize when entities are mathematically equivalent. Read information from graphs, tables, texts, or other sources. Order: Order and classify numbers, expressions, quantities, and shapes by common properties. Compute: Compute arithmetic operations with whole numbers, fractions, decimals, and integers using algorithmic procedures. Carry out straightforward algebraic manipulation.

Applying (Math):

The applying domain involves the application of mathematics in a range of situations. Problem solving is central to this domain. Students will need to select suitable operations,

strategies, and tools for solving problems. Formulate: Determine efficient/appropriate operations, strategies, and tools for solving problems. Implement: Implement suitable strategies and operations to produce solutions to problems. Represent: Represent data in tables or graphs; create equations, inequalities, geometric figures, or diagrams that model problem situations; and generate equivalent representations for a given mathematical entity or relationship.

Reasoning (Math):

Reasoning mathematically involves logical, systematic thinking. It includes intuitive and inductive reasoning based on patterns and regularities that can be used to arrive at solutions to problems. Evidence of reasoning processes can be found in the explaining or justifying of a solution method, or the making of valid inferences on the basis of information and evidence. Reasoning is required in analyzing or generalizing mathematical relationships. It contains skills like: Analyze: Analyze, describe, or use relationships among numbers, expressions, quantities, and shapes. Integrate: Link different elements of knowledge, related representations, and procedures. Generalize: Make statements that represent relationships in more general and more widely applicable terms. Justify: Provide mathematical arguments to support a strategy or solution.

Knowing (Science):

Items in this domain assess students' knowledge of facts, relationships, processes, concepts, and equipment. Accurate and broad-based factual knowledge forms a foundation that students can draw upon to successfully engage in the more complex cognitive activities essential to the scientific enterprise. It contains skills like: Recognize: Identify or state facts, relationships, and concepts; identify the characteristics or properties of specific organisms, materials, and processes; identify the appropriate uses for scientific equipment and procedures; and recognize and use scientific vocabulary, symbols, abbreviations, units, and scales. Describe: Describe or identify descriptions of properties, structures, and functions of organisms and materials, and relationships among organisms, materials, and processes and phenomena. Provide Examples: Provide or

identify examples of organisms, materials, and processes that possess certain specified characteristics; and clarify statements of facts or concepts with appropriate examples.

Applying (Science):

Items in this domain require students to engage in applying knowledge of scientific facts, relationships, processes, concepts, equipment, and methods in contexts likely to be common in the teaching and learning of science. It contains skills like: Compare/Contrast/Classify: Identify or describe similarities and differences between groups of organisms, materials, or processes; and distinguish, classify, or sort individual objects, materials, organisms, and processes based on characteristics and properties. Relate: Relate knowledge of an underlying science concept to an observed or inferred property, behavior, or use of objects, organisms, or materials. Interpret Models: Use a diagram or other model to demonstrate knowledge of science concepts, to illustrate a process, cycle, relationship, or system, or to find solutions to science problems. Interpret: Information Use knowledge of science concepts to interpret relevant textual, tabular, pictorial, and graphical information. Explain: Provide or identify an explanation for an observation or a natural phenomenon using a science concept or principle.

Reasoning (Science):

Items in this domain require students to engage in reasoning to analyse data and other information, draw conclusions, and extend their understandings to new situations. Scientific reasoning also encompasses developing hypotheses as well as designing scientific models and investigations. In contrast to the more direct applications of science facts and concepts exemplified in the applying domain, items in the reasoning domain may involve less common or more complicated contexts. Answering such items can involve more than one approach or strategy. It contains skills like: Predict: Formulate questions that can be answered by investigation and predict results of an investigation given information about the design; use scientific evidence and conceptual understanding to make predictions about the effects of changes in biological or physical conditions or about the outcome of a dynamic situation; and formulate testable assumptions based on conceptual understanding and knowledge from experience, observation, and/or analysis

of scientific information. Design: Develop models; plan investigations or procedures appropriate for answering scientific questions or testing hypotheses; describe or recognize the characteristics of well-designed investigations in terms of variables to be measured and controlled and cause-and-effect relationships; and design a plan that applies scientific principles and appropriate technologies to solve a problem. Evaluate: Evaluate alternative explanations; weigh advantages and disadvantages to make decisions about alternative processes and materials; evaluate models in terms of their merits and limitations; evaluate results of investigations with respect to sufficiency of data to support conclusions; and evaluate design plans in terms of criteria for success and constraints. Draw Conclusions: Make valid inferences on the basis of observations, evidence, and/or understanding of science concepts; and draw appropriate conclusions that address questions or hypotheses, and demonstrate understanding of cause and effect. Analyze: Identify the elements of a scientific problem and use relevant information, concepts, relationships, and data patterns to answer questions and solve problems. Synthesize: Answer questions that require consideration of a number of different factors or related concepts. Generalize: Make general conclusions that go beyond the experimental or given conditions; apply conclusions to new situations. Justify: Use evidence and science understanding to support the reasonableness of explanations, solutions to problems, and conclusions from investigations.

4.1.2.2 Content Domains

The following definitions were used for TIMSS Grade 4 content domains:

Number

Number provides the foundation of mathematics in primary school. The number content domain consists of three topic areas: Whole numbers, Expressions, simple equations and relationships, and Fractions and decimals.

Geometric Shapes and Measures

Geometry helps us visualize and understand the relationships between shapes and sizes. Measurement is the process of quantifying attributes of objects and phenomena (e.g., length and time). At the fourth grade, students should be able to use a ruler to measure length; perform calculations involving length, mass, volume, and time; calculate areas of shapes based on rectangles; calculate perimeters of polygons; and use cubes to determine volumes. Students should be able to identify the properties and characteristics of lines, angles, and a variety of two- and three-dimensional shapes. Spatial sense is integral to the study of geometry, and students will be asked to describe and draw a variety of geometric figures. They also should be able to analyze geometric relationships and use these relationships to draw conclusions about geometric objects.

Data Display

Students need to understand that graphs and charts help organize information or categories and provide a way to compare data. The data content domain consists of two topic areas: Reading and displaying data, Interpreting, combining, and comparing data. At the fourth grade, students should be able to read and create data displays. They should be able to make inferences from data displays and use data from one or more sources to answer questions of interest.

Life Science

Life science is represented by five topic areas: Characteristics and life processes of organisms, Life cycles, reproduction, and heredity, Organisms, environment, and their interactions, Ecosystems, and Human health.

Physical Science

Physical science content domain comprises: Classification and properties of matter and changes in matter, Forms of energy and energy transfer, and Forces and motion. Fourth grade students should have an understanding of physical states of matter (solid, liquid, and gas), as well as common changes in the state and form of matter; this forms a

foundation for the study of both chemistry and physics in the middle and upper grades. At this level, students also should know common forms and sources of energy and their practical uses, and understand basic concepts about light, sound, electricity, and magnetism. The study of forces and motion emphasizes an understanding of forces as they relate to movements students can observe, such as the effect of gravity or pushing and pulling.

Earth Science

Earth science is the study of Earth and its place in the Solar System, and at fourth grade focuses on the study of phenomena and processes that students can observe in their everyday lives. Topic areas included in this domain are generally considered to be important for students at the fourth grade to understand as they learn about the planet on which they live and its place in the Solar System: Earth’s physical characteristics, resources, and history, Earth’s weather and climates, and Earth in the Solar System. At this level, students should have some general knowledge about the structure and physical characteristics of Earth’s surface, and about the use of Earth’s most important resources. Students also should be able to describe some of Earth’s processes in terms of observable changes and understand the time frame over which such changes have occurred. Fourth grade students should also demonstrate some understanding about Earth’s place in the Solar System based on observations of patterns of change on Earth and in the sky.

4.1.3 The Complex Setup

The second setup mostly resembled the first one, except for two modifications visible in the generation query below:

“Create one item for the Polish TIMSS assessment. Do not write anything else than the content of the item. The item has to be in Polish. Adhere to the guidelines outlined below

Subject : {subject}

General Guidelines: {general_guidelines}

Content Domain : {content_domain}

Content Domain Guidelines: {content_domain_definitions_long[content_domain]}

Cognitive Domain : {cognitive_domain}
 Cognitive Domain Guidelines: {cognitive_domain_definitions[cognitive_domain + f' {subject[0]}']}
 Age appropriateness: Grade 4 of the elementary school
 Item format: {item_format_dict[item_type]}
 {hard_query_dict[subject[0]]}
 Additional comments: DO NOT GENERATE LINKS TO TABLES AND IMAGES, BUT PRESENT A DETAILED DESCRIPTION OF TABLES AND IMAGES IF YOU CHOOSE TO USE THEM IN YOUR ITEM (YOU DONT HAVE TO)"

The first modification related to the guided generation of specific item types – either constructed response (open-ended), or multiple-choice (closed-ended) items, achieved by adding the Item format condition:

"Item format: {item_format_dict[item_type]}"

This part of the prompt took the following form in the case of constructed response items:

"Constructed response. Constructed response items usually require students to give a numerical result³, provide a short explanation or description given in one or two phrases or sentences, or create a list"

or, in the case of the multiple-choice items:

"Multiple-choice. Multiple choice items have four response options, ONE OF THEM IS ALWAYS CORRECT AND THREE ARE INCORRECT. The incorrect options need to be plausible choices, demonstrating typical errors or misunderstandings. They need to be similar to the correct option in terms of length and format, but be undoubtedly incorrect."

The third prompt design focused on adjusting the difficulty towards a higher level. The choice to do this was motivated by the inspection of the items generated in the simple setup, which were deemed a little too simplistic. This modulation was achieved by adding either "MAKE THE ITEM CHALLENGING FOR THE STUDENT BUT DONT GO BEYOND THE LIMITS OF THE GUIDELINES", in the case of science items or "MAKE THE ITEM A BIT CHALLENGING FOR THE STUDENT BUT DONT GO BEYOND THE LIMITS OF

³ This prompt part is a bit unfortunate, as it applies only to mathematics items, but not to science items, however, it was used in generating both mathematics and science items.

THE GUIDELINES” in the case of math items. The difference between the two queries consisted of the added “A BIT” before the word “CHALLENGING” in the case of math items, without which the LLM generated items that were deemed too difficult for Grade 4 students. This technique of prompt engineering, often called “prompt augmentation”, serves different purposes, including to enhance the output quality (Hughes et al., 2024).

Apart from the differences outlined above, the complex setup did not differ in any other way from the simple setup.

4.1.4 The Final Choice of Items

Out of the generated items in two setups, the first items that achieved basic criteria of correctness (all response options are different, presence of a correct response option, lack of incorrect statements) were chosen to be included in the final set. Examples of excluded items are available in Annex 4. We have also excluded repetitive items, e.g. close or complete duplicates. In total, 64 items were picked from among the newly generated ones, half from the simple setup, and another half from the complex setup. In each of the setups, the items were split half/half to two different formats (MC and CR).

Additionally, 32 items were manually picked from among those that were rated favourably in the SMEs study. Apart from requiring high expert ratings, especially on the Usefulness, Correctness, and Difficulty criteria. The items selected also had to conform to the TIMSS guidelines with regard to item format and measured content and cognitive domain in a test booklet. These items came from all setups tested in the SMEs panel (zero-, one-, three-, and five-shot).

These actions led to a set of 96 LLM-generated items grouped in three setups (experimental conditions):

- 32 items from *simple* condition,
- 32 items from *complex* condition,
- 32 items from *teachers’* condition.

4.1.5 Generating Images

In cases where the LLM-generated items with image descriptions, or ones that directly implied the need for an image, the images were generated with the help of the OpenAI interface, and the gpt-4 model. The images were generated in an iterative way until a satisfactory image was achieved. In cases where the images were of geometrical nature, the LLM was prompted to use its coding abilities to generate them using python and matplotlib.

4.2 Test preparation

4.2.1 Test plan (incomplete block design)

In order to test more items, we have decided to use a missing-be-design test plan, meaning that not all participants respond to every item used in a test. Instead, the test is divided into subsets of items (e.g., blocks or forms), and these subsets are administered to different groups of test-takers. This results in some items "missing" for certain participants by design, rather than due to random non-response or other issues. This test plan allows the testing of a large number of items, while keeping the test length for one testee within reasonable limits (Pokropek, 2011).

To keep our test plan as close as possible to the original TIMSS, we constructed sixteen item blocks containing different items with mathematics and science parts of the assessment. Each block contained a similar number of items as in original TIMSS (around 13), at least half of which were automatically generated. The remaining items came from the pool of TIMSS Grade 4 restricted-use items. Even-numbered blocks consisted of real TIMSS Grade 4 items, while odd-numbered blocks consisted of automatically generated items⁴. In this design we used 192 items (half generated, half real), which, given that we aimed to recruit around 1000 students, yielded around 250 students per item - a sufficient number to estimate IRT models (cf. Pokropek, 2011). While the design is not fully balanced, it facilitates complete linkage between items, comprising booklets that blend real and generated items, as well as introducing booklets composed exclusively of real

⁴ The items from two of the three setups were mixed in roughly equal proportions in one block.

items alongside those that will be generated. The overall booklet design for this assessment is displayed in Figure 10.

Figure 10

Overview of Test Booklets

	Part 1		Part 2	
	A	B	C	D
Book id 1a	1	2	3	4
Book id 1b	4	3	2	1
Book id 2a	5	6	7	8
Book id 2b	8	7	6	5
Book id 3a	9	10	11	12
Book id 3b	12	11	10	9
Book id 4a	13	14	15	16
Book id 4b	16	15	14	13
Book id 5a	1	5	9	13
Book id 5b	13	9	5	1
Book id 6a	2	6	10	14

Book id 6b	14	10	6	2
Book id 7a	15	11	7	3
Book id 7b	3	7	11	15
Book id 8a	4	8	12	16
Book id 8b	16	12	8	4

4.2.2 Assessment contents

Each booklet consisted of four blocks, two from mathematics, two from science, to a total number of sixteen booklets in the assessment. In order to counter fatigue effects, we counterbalanced the order of the domains, with A booklets starting with mathematics and B booklets starting with science. Items to each of the booklets were selected to fulfill guidelines of TIMSS assessment framework with regard to proportion of items' content and cognitive domain, format (50% MC and 50% CR) and presence of additional graphical information, e.g. image (50% with image) in one booklet. TIMSS items were allocated to the booklets in order to fulfill these criteria whenever it was possible. Sometimes, we did not have access to enough TIMSS items to meet all the assessment framework criteria, because we could only use TIMSS items from 2011, 2015, and 2019 cycles. While we were preparing the assessment TIMSS 2023 items were not available, moreover, Poland has participated in TIMSS only from 2011, hence older TIMSS items do not have an official Polish version, which we considered a serious limitation and consequently decided to constrain the item selection to TIMSS cycles in which official Polish items were available. What is more, we were not able to use TIMSS graphical-response items due to limitations of our testing platform, which further restricted the choice of TIMSS CR items.

The test procedure was kept as close as possible to the original TIMSS. students first read instructions along with the interviewer, then completed one booklet for 36 minutes, had a 10-minute break, and then completed a second booklet (again 36

minutes), hence the whole study lasted approximately two 45-minute lessons. At the end of the second booklet, students responded to a short questionnaire, which summarized in Table 13. The study was conducted with written approval from the Institute’s ethical research committee.

Table 13

Summary of Questionnaire Items

Measured construct	Number of items (format)
Gender	1 (dichotomous)
Test self-reported effort	1 (4-point rating scale)
Materials used during the test (e.g. telephone)	4 (dichotomous)
Technical problems survey	1 (5 options to choose)
Interviewer provided instructions	1 (5-point rating scale)
Number of books at home	1 (5 options to choose)
Mouse use behaviour	1 (4 options to choose)

Note. The question about the number of books at home was taken from the original TIMSS questionnaire as a basic measure of students’ family socioeconomic status.

4.2.3 Paradata collection

Paradata was collected using the *logdataLimeSurvey* version 1.2 Java Script applet installed on the *LimeSurvey* web surveying platform. A description of the technical details and list of collected events is available in the applet’s web page <https://github.com/tzoltak/logdataLimeSurvey>. Participants were informed about the paradata collection (response times and cursor movements) in the debriefing instruction. Response times data is analysed in this report, while cursor moves and other paradata is not. These data will be analysed in other publications.

4.3 Fieldwork preparation, procedures and timeline

4.3.1 School recruitment and remuneration

The fieldwork procedures were designed to standardise 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.

The recruitment of schools was carried out by a recruited contractor - a specialised external research company. According to Polish law and IFiS PAN internal regulations, a tender was organized to select a company to conduct the fieldwork operations. The Market Research World Anna Sojka company was chosen among the total of six applicants on the basis of the price and other conditions offered.

The contractor contacted schools via telephone and personal interactions. Whenever a school agreed to participate in the study, all available students from all classes were qualified to participate. 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. The final decision on student-level exclusions was made by the school principal 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.

Entire classes participating in the study received small gifts from the research company in the form of board games or book sets for the total value of around 200 PLN (ca. 50 €) per classroom. The classrooms could choose what games or books exactly they would be given. Individual students were not rewarded for participation. Schools received a certificate of attendance and a tailored school report, presenting school results in comparison to other schools participating in the study, as well as a list of (anonymised) individual student results. These products were requested by schools.

4.3.2 Testing room preparation

Before data collection started, all participating interviewers took part in a two-hour 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.

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 the internet browser type and version was included in the paradata collected by the testing platform.

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.

4.3.3 Assessment session procedures

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 students.
- 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 and science tasks and a questionnaire at the end). The instructions also informed students 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.

Table 14 presents the flow of testing sessions, with all the assessment parts and its estimated time of duration. The order of the mathematics and science parts was counterbalanced across test takers and was randomly assigned to a given student.

Table 14

Overview of Testing Session Flow

Activity - version A	Activity - version B	Time for participant
1. Preparing participants for the study: logging in, going through the introductory part, written survey instructions, etc.	1. Preparing participants for the study: logging in, going through the introductory part, written survey instructions, etc.	5 minutes
2. First part of assessment (mathematics)	2. First part of assessment (science)	35 minutes
3. Break	3. Break	10 minutes
4. Second part of assessment (science)	4. Second part of assessment (mathematics)	35 minutes
5. Questionnaire	5. Questionnaire	2 minutes

Note. Column “Time for participant” presents approximate times. Version A and B differed in the order of assessment domains: version A displayed mathematics items first, while version B started from science items.

4.3.4 Pilot administration

Before the main data collection started, the assessment procedures were tested in two schools. Some technical problems with the testing platform were identified and corrected by the researcher team.

4.3.5 Data collection timeline

The data collection took place between early September and mid November 2024 and comprised almost 1400 Grade 4 students from 36 schools. The schools were all located in the south of Poland.

Applying a number of exclusion criteria reduced the final sample size to 1200 students. This was done in order to exclude participants who did not conform to the assessment instructions and rules. The excluded observations **were not used in any of the analyses presented in the report**. The set of criteria for excluding observations comprised:

- **Drop Off.** The survey was not completed.
- **Speeding.** Survey completion time below 26 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. The main reason for that was the human error by the interviewer. 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.

Table 15 displays the complete fieldwork timeline, with dates of all assessment sessions, assessed students in each day and number of excluded students and reasons of exclusion.

Table 15
Fieldwork Data Collection Timeline

Day	N. of sessions	Number of all records	Number of accepted records	Number of excluded records					
				Overall	Not completed	Speeding	Left the browser tab	Same login used more than one time	Started too late
09.09	3	54	53	1	0	0	0	1	0
13.09	1	24	19	5	2	1	3	0	0
20.09	1	11	9	2	1	1	0	0	1
23.09	2	33	26	7	4	4	3	0	2
24.09	6	86	60	26	26	23	0	0	1
25.09	1	25	24	1	0	0	1	0	0
26.09	2	48	36	12	2	1	9	0	1
27.09	2	39	32	7	2	2	4	0	0
30.09	3	66	65	1	0	1	0	0	0
02.10	5	64	60	4	2	2	1	1	2
03.10	2	27	20	7	2	1	5	0	0
04.10	6	87	81	6	1	1	0	5	0
08.10	10	148	107	41	12	5	24	0	4
09.10	2	20	19	1	1	0	0	0	0
10.10	1	11	11	0	0	0	0	0	0
11.10	1	13	13	0	0	0	0	0	0
15.10	1	14	14	0	0	0	0	0	0
16.10	4	78	74	4	3	3	1	1	0
17.10	1	18	17	1	1	0	0	0	0
18.10	4	74	46	28	0	0	0	0	28
21.10	2	33	29	4	4	0	0	0	0
22.10	4	39	38	1	0	0	0	0	1
23.10	4	44	38	6	1	1	0	4	1
24.10	3	34	32	2	0	0	0	2	0
25.10	2	24	24	0	0	0	0	0	0
28.10	2	18	18	0	0	0	0	0	0
29.10	2	47	42	5	1	1	3	0	1
30.10	3	40	40	0	0	0	0	0	0
05.11	2	41	41	0	0	0	0	0	0
06.11	3	50	50	0	0	0	0	0	0

Day	N. of sessions	Number of all records	Number of accepted records	Number of excluded records					
				Overall	Not completed	Speeding	Left the browser tab	Same login used more than one time	Started too late
07.11	3	47	47	0	0	0	0	0	0
14.11	2	40	16	24	12	11	0	0	13
Overall	90	1397	1201	196	77	58	54	14	55

4.3.6 Sample composition

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 16 and 17 display sample characteristics with regard to these two variables. The final sample was perfectly balanced with regard to gender, but there were some slight diversions from the target in case of school location size.

Table 16
Sample Gender Distribution

Gender	Number	Percent	Target
Male	604	50.3%	50%
Female	597	49.7%	50%
Overall	1201	100.0%	100%

Table 17*Sample Distribution by School Location*

School location size	Number	Percent	Target
<5k	161	13.4%	35%
5k-100k	770	64.1%	40%
>100k	270	22.5%	25%
Overall	1201	100.0%	100%

Note: <5k - school location size smaller than 5,000 inhabitants; 5k-100k - school location size between 5,000 and 100,000 inhabitants; > 100k - school location size bigger than 100,000 inhabitants.

4.3.7 Booklet Coverage

In order to present all items a similar number of times the sixteen test booklets needed to be presented also a similar number of times. Table 18 displays that this was indeed achieved and a similar number of students worked with any given test version.

Table 18*Booklet distribution in the analyzed sample*

Booklet	Number	Percent
1A	80	6.7%
1B	70	5.8%
2A	81	6.7%
2B	75	6.2%
3A	78	6.5%
3B	67	5.6%
4A	85	7.1%

Booklet	Number	Percent
4B	74	6.2%
5A	77	6.4%
5B	71	5.9%
6A	77	6.4%
6B	71	5.9%
7A	68	5.7%
7B	67	5.6%
8A	85	7.1%
8B	75	6.2%
Overall	1201	100.0%

Note. Chi-squared test against the uniform distribution: $\chi^2(15) = 6.83, p = 0.962$

4.3.8 Item coding

Multiple-choice items responses were recoded to 0-1 on the basis of the TIMSS scoring rules in the case of original TIMSS items and on the basis of LLM scoring rules reviewed by the research team members in the case of LLM-generated items. Constructed-response items were scored on the similar basis: students' responses were coded by team members on the basis of TIMSS scoring guidelines and scoring guidelines generated by LLM, explained in the section 4.3.9 below. The latter were used in the case of LLM-generated items after being reviewed by the team members. We have also used the LLMs to score the items, which yielded a high correlation with human coding. However, this analysis is beyond the scope of this report and will be presented elsewhere.

4.3.9 Generating scoring guidelines

To generate scoring guidelines, we first manually selected three item-scoring key pairs for mathematics and three for science (six in total) from the English TIMSS manuals, ensuring coverage of different item types. The original items, scoring keys, and their translations are presented in Annex 5. These materials were extracted using the GPT-4o

image API endpoint, with the following prompts applied consecutively for item and scoring key extraction:

Exercise extraction prompt:

```
"Extract exercises text from this TIMSS exam page image in a text format.  
Whenever tables appear, encode them using markdown alongside the exercise  
text. Whenever images appear, describe what they portray in the format [IMAGE:  
image_description] and add the description to the text in the proper place.  
Make sure to extract all exercises visible, and do not write anything else!"
```

Scoring key extraction prompt (with the previously extracted item text appended):

```
"Extract the scoring key of the item presented below from this TIMSS exam manual  
page image in a text format.  
Extract the key in a narrative format, explaining in detail what should be  
taken into account when grading this image. Disregard answer codes and only  
focus on the information that is essential to either award or deduct points.  
Do not preface with any introduction, return only raw scoring instructions.  
Focus on both the correct and incorrect response specifications. Do not repeat  
the item text.  
Item: {item}  
Scoring instructions:"
```

The extracted item-scoring key pairs were then translated using the GPT-4o API endpoint, with the same prompt-based approach. All hyperparameters were kept identical to those used in the previous generation step: $top_p = 1$, $n = 1$, $presence_penalty = 0$, $frequency_penalty = 0$. The max_tokens parameter was left unspecified, allowing the model to generate output up to the maximum length permitted by the context window. The only exception was the $temperature$ parameter, which was set to a fixed value, as output variability was not required for this task.

```
"Translate the following to Polish. Do not write anything else.  
###  
Text to translate:  
{item}  
###"
```

Translation:

To generate scoring keys for the Polish LLM-generated items we had created, we again used the GPT-4o API endpoint. The hyperparameters were kept consistent with the previous generation process: $top_p = 1$, $n = 1$, $presence_penalty = 0$, $frequency_penalty = 0$. The max_tokens parameter was left unspecified, allowing the model to generate output up to the maximum length allowed by the context window. As before, the only exception was the $temperature$ parameter, which was fixed, since output variability was not required for this task. Given that the items were in Polish, we used the following Polish-language prompt to prevent the model from mixing different languages:

Original prompt:

"Stwórz klucz odpowiedzi do sprawdzenia następujących zadań. Nie pisz nic poza kluczem. W kluczu uwzględnij zarówno poprawne odpowiedzi, jak i odpowiedzi błędne w formacie Prawidłowa odpowiedź: [instrukcje], Nieprawidłowa odpowiedź: [instrukcje].

Jeśli w tekście zadania nie wspomniano o potrzebie wyjaśnienia odpowiedzi, lub zapisania obliczeń, nie odejmuj punktów za brak tych elementów."

Translation:

"Create an answer key for grading the following tasks. Do not write anything except the key itself. In the key, include both correct and incorrect answers in the format:

Correct answer: [instructions],

Incorrect answer: [instructions].

If the task text does not mention the need to explain the answer or show calculations, do not deduct points for the absence of these elements."

This prompt was then extended with a continuation that included the previously extracted item-scoring key pairs (for either mathematics or science), the maximum number of points available for each item, and any additional information required for interpreting or completing the task. This supplementary information was added in cases where the item text alone was insufficient to complete the task, for example, when the

item required calculating the area of a triangle shown in an image, but the textual description did not fully specify the necessary dimensions.

Original Continuation:

```
"""###  
ZADANIE:  
{row['item_PL']}  
KLUCZ ODPOWIEDZI:  
{row['instruction_PL']}  
Maksymalna ilość punktów: {str(points)}  
[[Dodatkowe informacje potrzebne do wygenerowania klucza: {additional_info}]]"
```

Translation:

```
"""### EXERCISE:  
{row['item_PL']}  
ANSWER KEY:  
{row['instruction_PL']}  
Maximum number of points: {str(points)}  
[[Additional information needed to generate the key: {additional_info}]]"
```

And finally, the item was appended in the following manner:

Original:

```
"""###  
ZADANIE:  
{item}  
KLUCZ ODPOWIEDZI:  
"
```

Translation:

```
"""###  
EXERCISE:  
{item}  
ANSWER KEY:  
"
```

The generated answer keys were validated for correctness and refined by the team members in the process of scoring the items.

4.4 Psychometric modeling

4.4.1 Excluded items

Two items were excluded from the psychometric analyses:

- S100006 (LLM item) - because of having no variance in scores (only incorrect answers),
- M051134 (TIMSS item) - because of the technical error which occurred during its implementation at the testing platform.

4.4.2 Model estimation and model fit

IRT 2PL models were estimated using the *mirt* R package, version 1.42. (Chalmers, 2012). There were three unidimensional models estimated separately for each content domain:

- 1) model including all items (both original TIMSS and LLM-generated),
- 2) model including only TIMSS items,
- 3) model including only LLM-generated items.

Additionally, we have specified a rival model for each domain: a two-dimensional model with TIMSS items and LLM-generated items assigned to separate, correlated dimensions.

Table 19 contains values of fit indices for each model. We have also calculated the likelihood test (LRT) of the two-dimensional model pitted against the unidimensional model including all the items, as well as a latent correlation between the TIMSS and LLM-generated dimensions in the two-dimensional models.

Table 19*Model Fit Statistics for IRT Models*

	logLik	AIC	BIC	SABIC
All math items - unidimensional	-14,936.2	30,260.4	31,248.1	30,631.9
All math items - separate TIMSS and LLM dimensions	-14,914.0	30,218.1	31,210.8	30,591.4
All science items - unidimensional	-15,064.8	30,523.7	31,526.6	30,900.8
All science items - separate TIMSS and LLM dimensions	-15,058.2	30,512.3	31,520.3	30,891.4

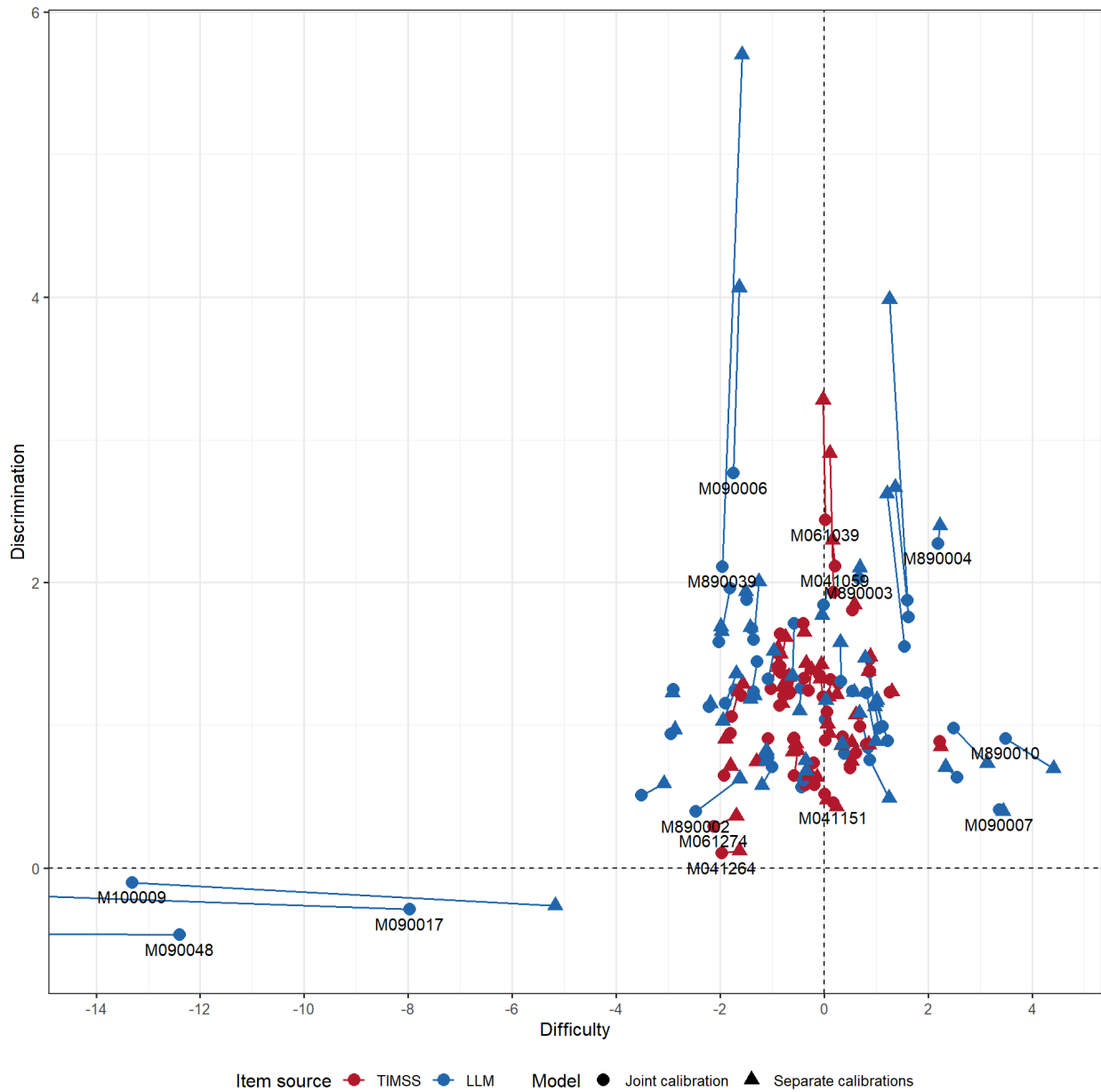
Note. logLik - log-likelihood, AIC - Akaike's Information Criterion, BIC - Bayesian Information Criterion, SABIC - Sample-adjusted Bayesian Criterion. Better fit in bold.

The results of model fit statistics, as well as LRT tests favoured the two-dimensional model over one-dimensional model, both in case of mathematics ($\chi^2(1) = 44.4$, $p = 0.000$) and science ($\chi^2(1) = 13.4$, $p = 0.000$). This result may be interpreted as TIMSS and LLM-generated items possess different characteristics and measure two distinct concepts. However, this difference is in practice negligible, as the two dimensions are strongly correlated – 0.88 in case of mathematics (with 95% CI limits for this IRT model parameter being 0.83 and 0.92) and 0.90 in case of science (with a 95% CI ranging from 0.85 to 0.96). Hence, it can be concluded that both TIMSS and LLM-generated items measure essentially the same construct.

As we can see in Figures 11 and 12 the two-dimensional structure may be an effect of some low-quality items present among the LLM-generated items. As it can be seen on these Figures as well, the joint calibration of IRT model, i.e. using both TIMSS and LLM-generated items to estimate IRT model parameters helps to obtain more reasonable parameters for some of the LLM-generated items. In the case of mathematics, such items are, e.g., M090006 or M890039. Problems with estimating difficulty parameters in case of items such as M100009, M090048 or M090017 stems from their negative discrimination.

Figure 11

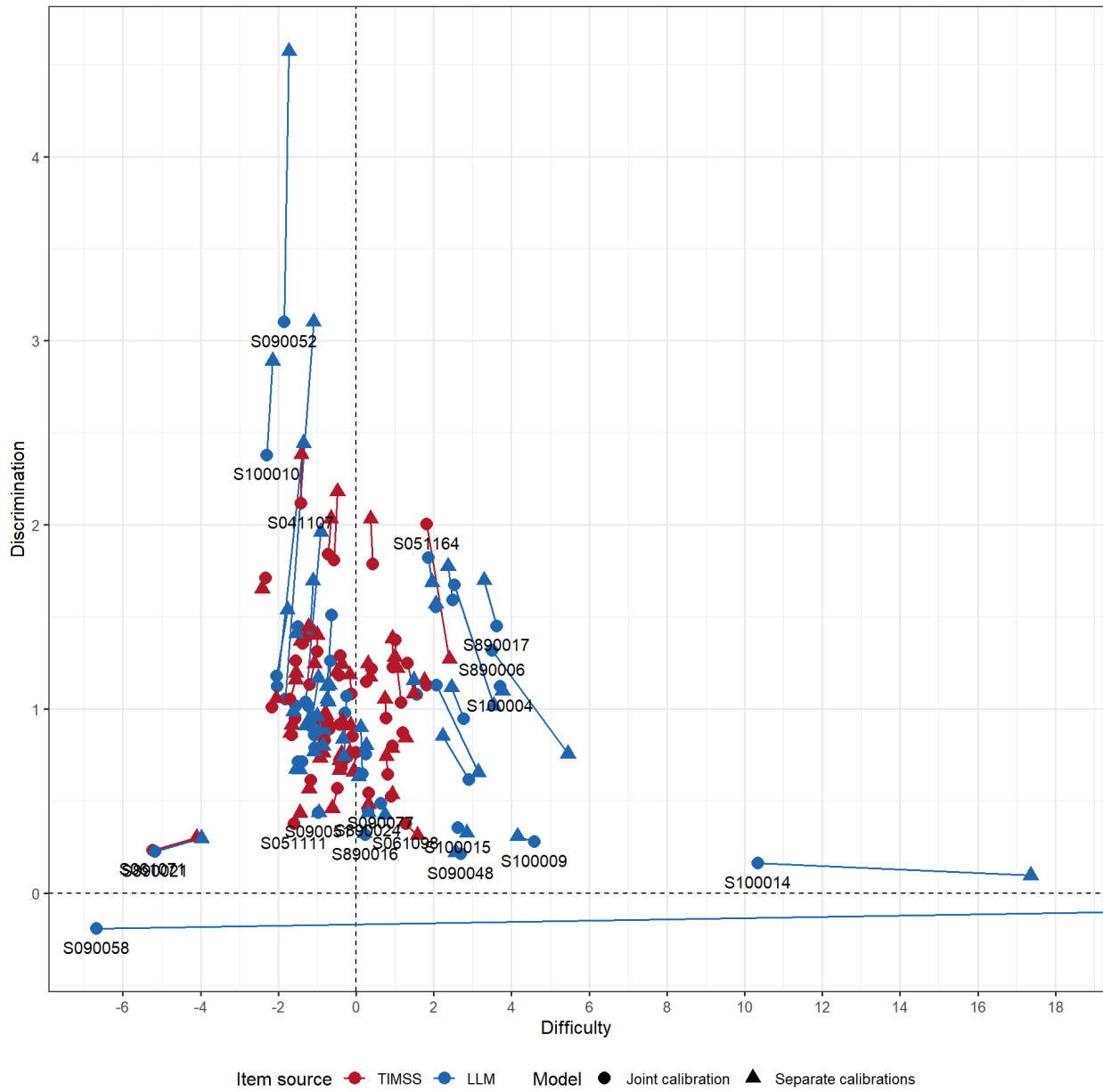
Difference between Joint and Separate IRT model calibration for Maths Items



In the case of science, such items are S090058 or S100014 (low discrimination causing problems in difficulty estimation) and S090052 or S100010 (joint calibration helps for unrealistic discrimination values).

Figure 12

Difference between Joint and Separate IRT model calibration for Science Items



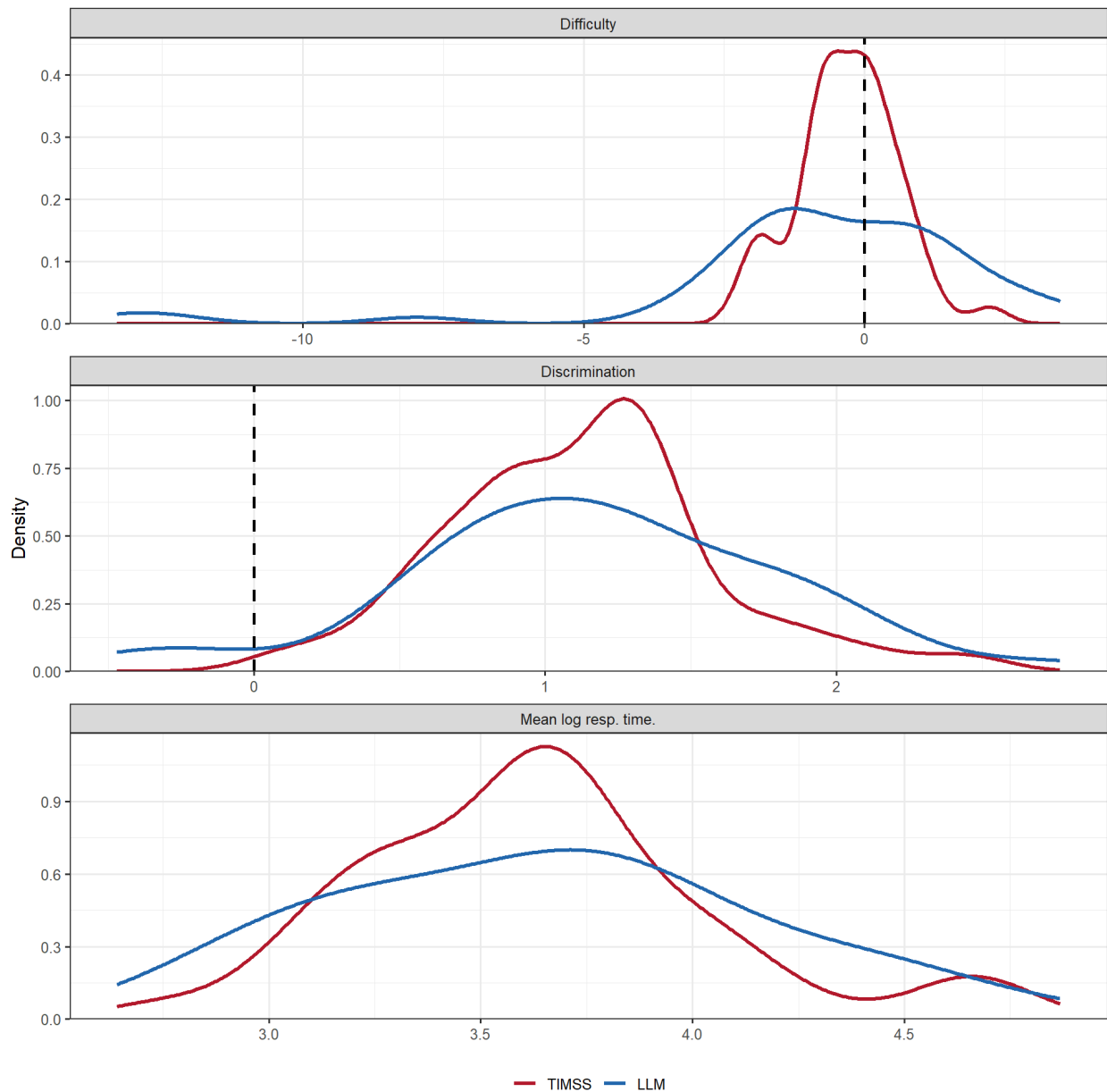
4.4.3 Item parameters of TIMSS and LLM-generated items

Already a first look at Figures 11 and 12 suggest problems with item parameters in case of some LLM-generated items. Both in mathematics and in science, TIMSS items almost never go beyond the range of $(-2, + 2)$ in case of difficulty, while this happens for a significant group of LLM-generated items. In mathematics, there are several items too difficult (difficulty parameter > 2) and several too easy (difficulty parameter $< - 2$), while in science there is a group of too difficult items. Problem of low discrimination (< 0.5) is not very common, but mainly relates to LLM-generated items, which, unlike TIMSS items, also note a few cases of negative discrimination parameter - a sign that probability of solving item correctly decreases with the increase of estimated student abilities.

Figures 13 and 14 present the distribution of item parameters for mathematics and science items, respectively. These parameters come from the joint calibration (this applies to all of the similar figures presented below).

Figure 13

Distribution of Item Parameters for Mathematics



The comparison between the distributions of difficulty parameters (uppermost panel) of TIMSS and LLM-generated items corroborates the above conclusions - while difficulty of TIMSS items is well within the range of $(-2, + 2)$ with the highest density over mean level of abilities, the LLM-generated items present a much more platykurtic distribution, with a lot of item parameters outside the expected range. It means that the difficulty of some of the LLM-generated items is not aligned to the tested sample and that

these items contribute to the precision of measurement only in the extremes of the range of abilities, where there are very few actual test-takers.

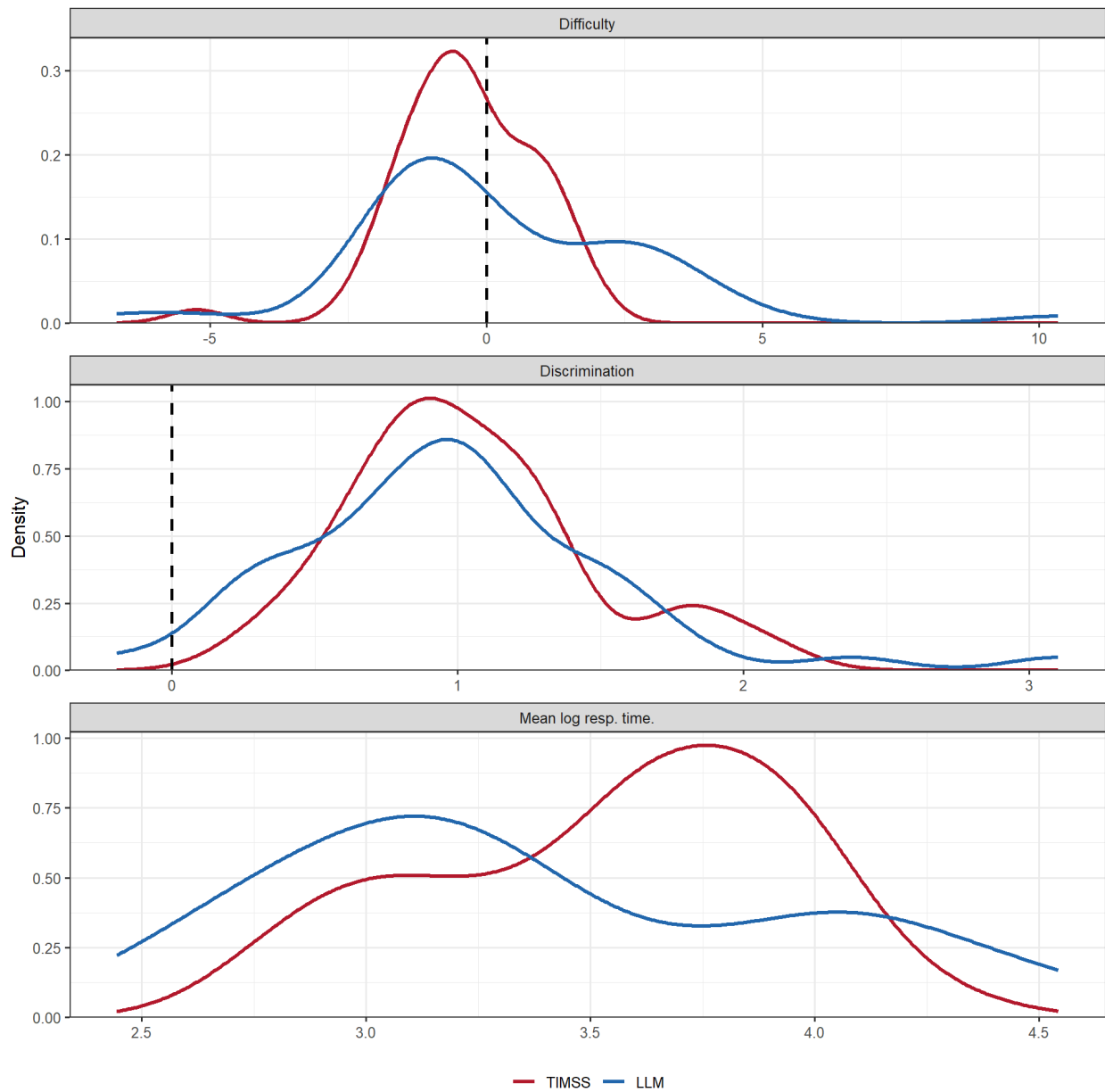
The comparison for discrimination parameters shows that these values follow a very similar distribution in case of TIMSS and LLM items. However, there are slightly more extreme discrimination parameters in case of LLM-generated items, including both very low or even negative parameters, as well as unrealistically high values, exceeding 2. These very high values of the discrimination parameter also signal problems, such as homogenous (item is too easy or too difficult) or overly polarised (item responses follow a bimodal distribution) response patterns or item redundancy, e.g. due to high correlations with other items.

Finally, the comparison of the distributions for item response times (logged) shows that LLM-generated items follow a more platykurtic distribution than TIMSS items, probably a consequence of a much wider range of item difficulties in case of LLM-generated items⁵. Logged item response times are usually positively correlated with item difficulty (on the item level), but both item (e.g. item discrimination) and participant (e.g. gender, item-solving strategies, motivation) characteristics moderate this relationship on the testee level (Chae et al., 2019; Embretson, 2021; Verbic & Tomic, 2009).

⁵ Encouraged by one of the reviewers we have analysed the correlation between item difficulty and response time, comparing TIMSS and LLM items in each of the formats (MC and CR). The results point to a positive correlation for both MC and CR TIMSS items in mathematics, but not for LLM items in this domain. The results for science are more nuanced - we again observe the positive correlation for MC TIMSS but not LLM items. In case of CR items, the situation is reversed - no correlation for TIMSS CR items, but a positive correlation for CR LLM items. We treat these results as tentative, as the power of this analysis is low (only around 20 items were compared in each of the cells).

Figure 14

Distribution of Item Parameters for Science



Comparison of key item parameter distributions between original TIMSS and LLM-generated items in the case of science leads to very similar conclusions as in the case of mathematics. Again, difficulty parameters for LLM-generated items follow a much more platykurtic distribution, with more very easy or very difficult items and fewer average items. Distribution of discrimination parameters is similar, but LLM-generated items have more items of low discrimination. TIMSS items show a higher mean logged response time

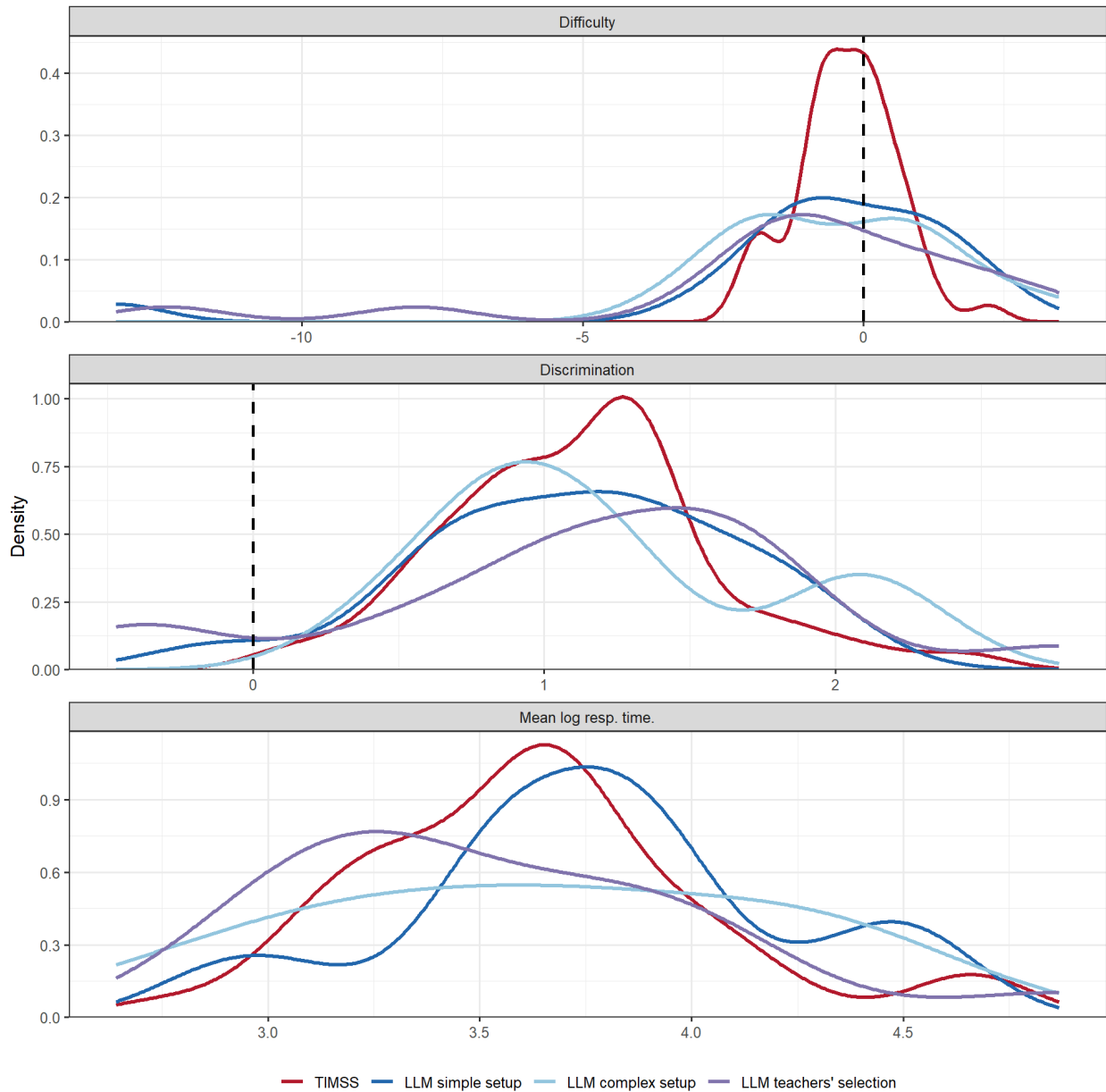
than LLM-generated items, which may tap into item-solving processes - students, confronted with overly difficult items may resort to rapid guessing (e.g. Lindner et al., 2019).

Comparison of Item Parameters between Generational Setups

Figures 15 and 16 go more in depth in analysing the distributions of item parameters, presenting separately the distributions for three LLM setups.

Figure 15

Distribution of Item Parameters for Mathematics: Comparison of LLM Setups



All three item-generation setups yielded similar item difficulty distributions, with the complex setup characterised by a larger group of difficult items than other setups, but also with more very easy items. It seems that this setup, tailored to generate more difficult

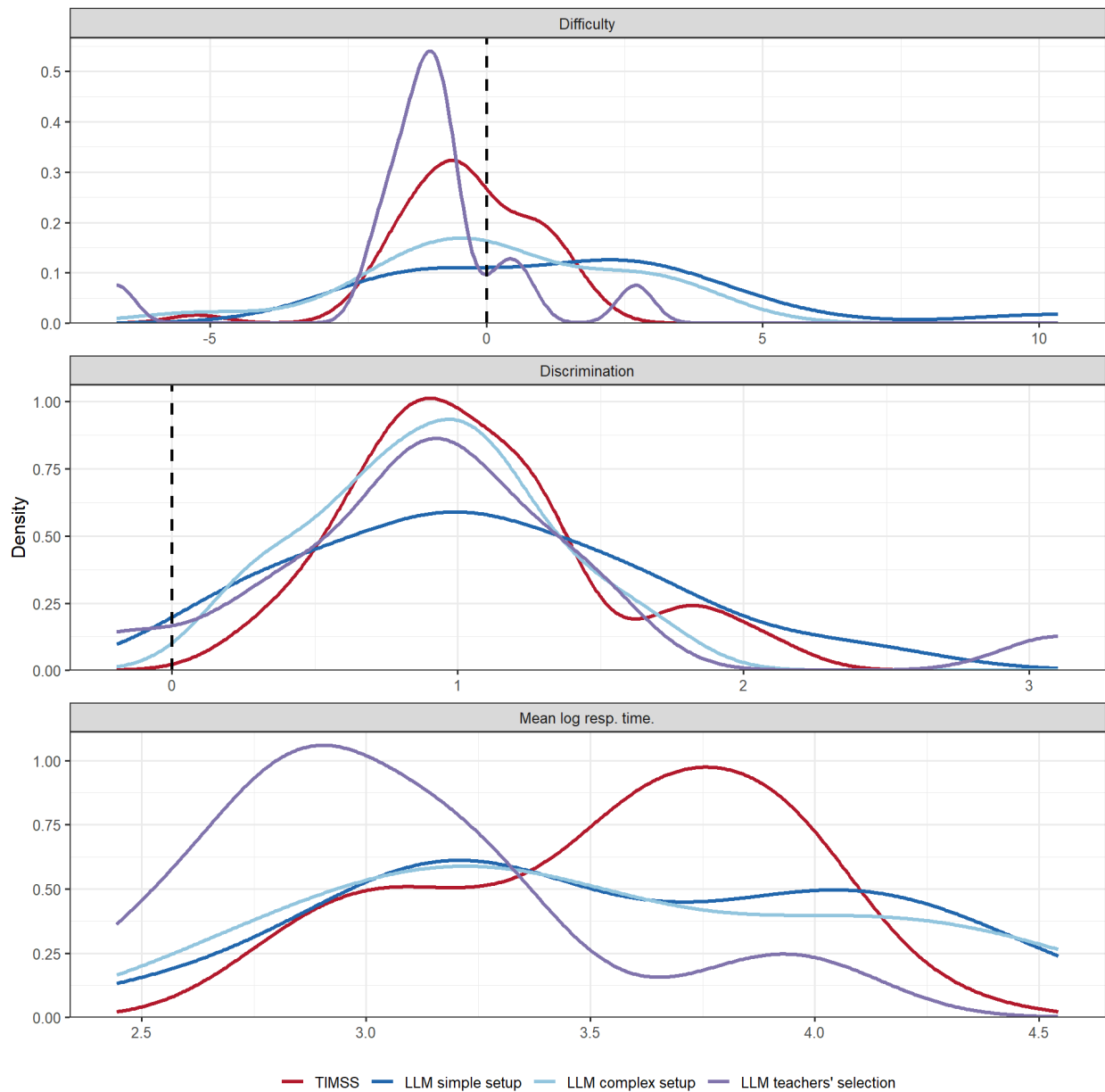
items, led to a bimodal distribution of item difficulties, with more demanding but also more overly easy items than the other two setups.

This bimodality was also observed in case of discrimination parameters distribution for this setup. LLM-simple setup yielded distribution that was the closest to the original TIMSS item discrimination distribution, while LLM-teachers' selection items were characterised by a larger group of items with very high discrimination (but within the reasonable range of 1.5-2.0). This group of LLM-generated items also noted the lowest item reaction times, while LLM-simple setup items had the longest.

These differences were further analysed with the use of non-parametric analogue of analysis of variance (ANOVA) - the Kruskal-Wallis rank sum test. We have resorted to using this test because the distributions of item parameters were often far from the assumed normality. The Kruskal-Wallis test did not yield any statistically significant results, not allowing to reject the null hypothesis of no cross-groups differences in central tendency. Exact test values are presented in Table S.1 in the Supplement.

Figure 16

Distribution of Item Parameters for Science: Comparison of LLM Setups



For science items, the LLM-generated items selected by the teacher contained significantly more easy and fewer difficult items than the TIMSS items. The other two LLM setups yielded items with a very broad range of item difficulties, ranging from very easy to very difficult items, with much fewer items of average difficulty (-0.5, 0.5) than TIMSS and LLM teachers' selection groups.

LLM-complex and LLM-teachers' selection setups yielded very similar item discrimination distributions to the original TIMSS items with the noteworthy over-presence of items with low discrimination parameter (< 0.5). On the other hand, LLM-simple setup yielded a more platykurtic distribution than other groups with more items of high discrimination (> 1.0), but also a larger group of items with extreme and unwanted discrimination parameters (below 0.5 or above 2.0).

Mean logged response times were the lowest in the LLM-teachers' selection setup, while they were the highest in the TIMSS items - probably a consequence of lower item difficulty in the LLM--teachers' selection setup.

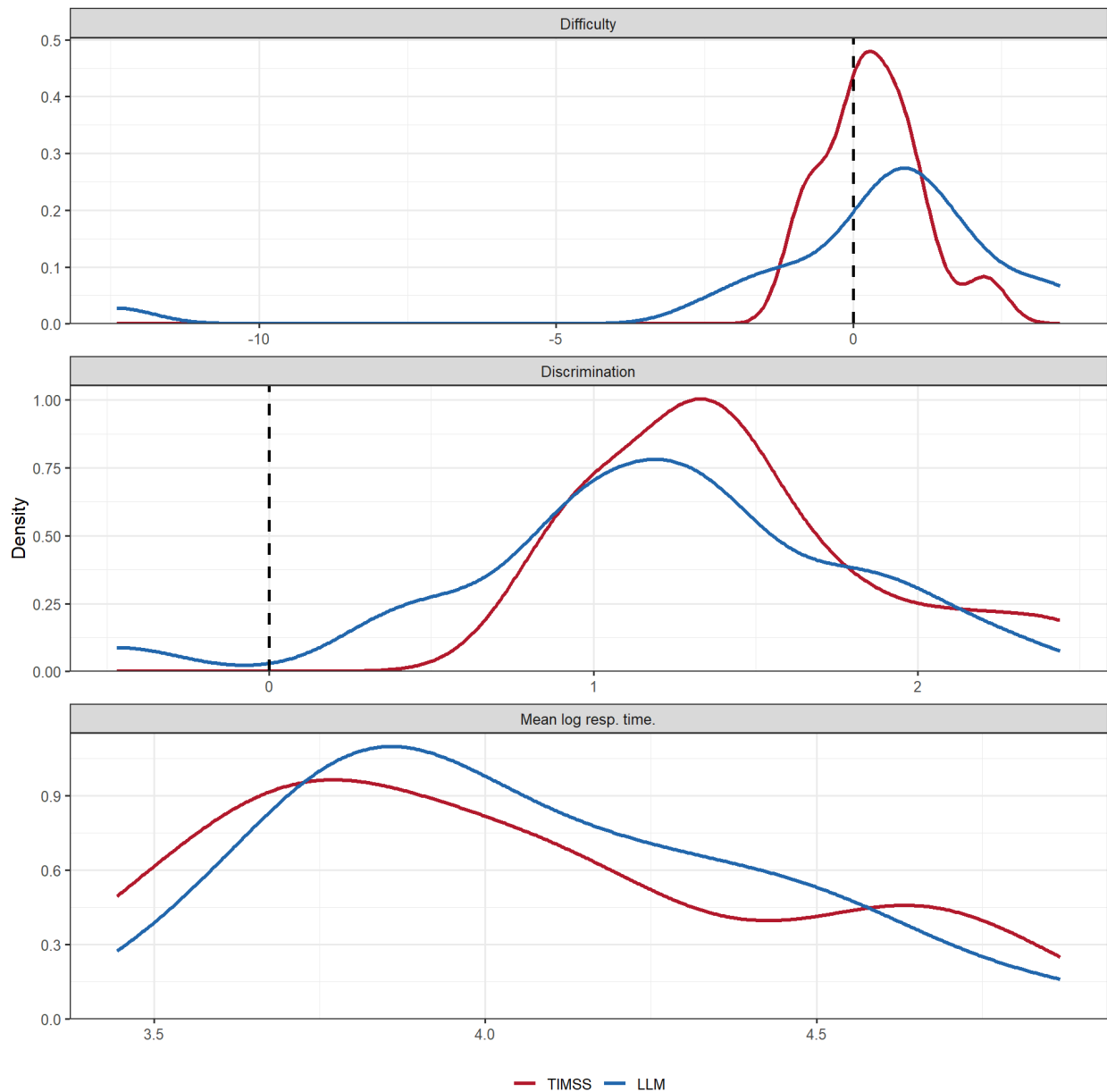
These differences were quantified again with the use of Kruskal-Wallis rank sum test that showed significant cross-group differences for difficulty ($\chi^2(3) = 9.21, p = 0.027$) and response time ($\chi^2(3) = 12.41, p = 0.006$) parameters. Exact test values are presented in Table S.2 in the Supplement. In order to qualify these differences, we conducted a series of post-hoc comparisons, to which we used a Wilcoxon rank sum exact test. These tests confirmed that TIMSS science items were easier than LLM-simple and LLM-teachers' selection items and that LLM-simple and LLM-complex were both more difficult than LLM-teachers' selection. Similar differences were observed with regard to mean response times: TIMSS multiple-choice mathematics items were also characterised by longer response times. A detailed presentation of these analyses is in Table S.3 in the Supplement.

Comparison of Item Parameters between Item Formats

We also decided to analyse item parameter distributions between items differing in formats, comparing multiple-choice (MC) with constructed-response (CR) items. These results are presented in Figures 17 and 19.

Figure 17

Distribution of Item Parameters for Mathematics: Constructed-Response Items



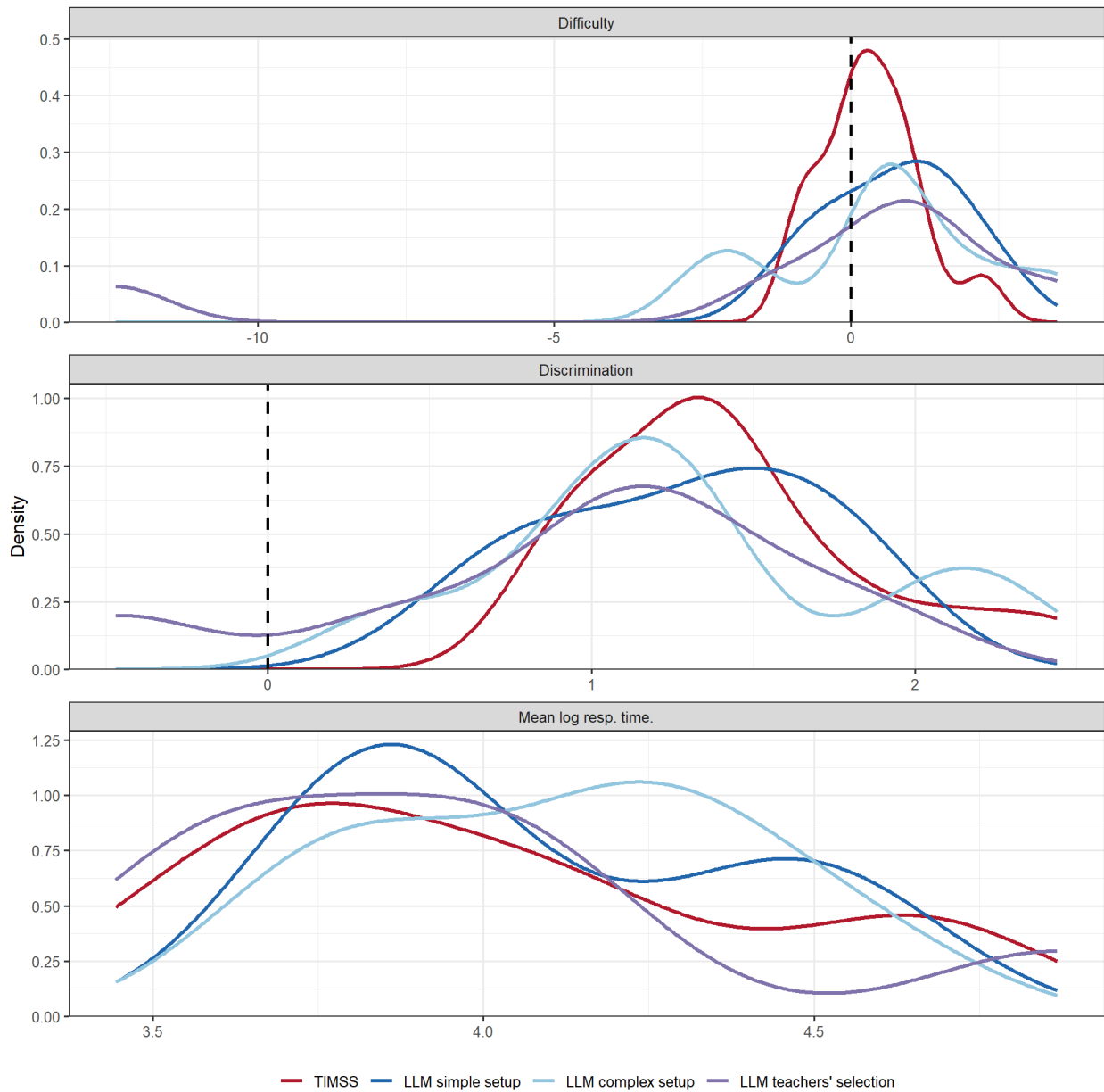
The difficulty parameters distribution of TIMSS items was leptokurtic and concentrated around the mean difference (0), while the distribution of LLM-generated item parameters exhibited more easy and difficult items. The distribution for discrimination parameters was similar in TIMSS and LLM-generated items, with a notable surplus of items with low (< 0.5) discrimination parameter in case of LLM-generated items. The distributions for mean logged response times differ slightly between TIMSS and LLM

items with the former yielding more items with very fast and very long logged response times and the latter with more items with average logged response times.

Figures 18 and 20 show these distributions in more detail by comparing the LLM-generating setups with TIMSS items and with each other.

Figure 18

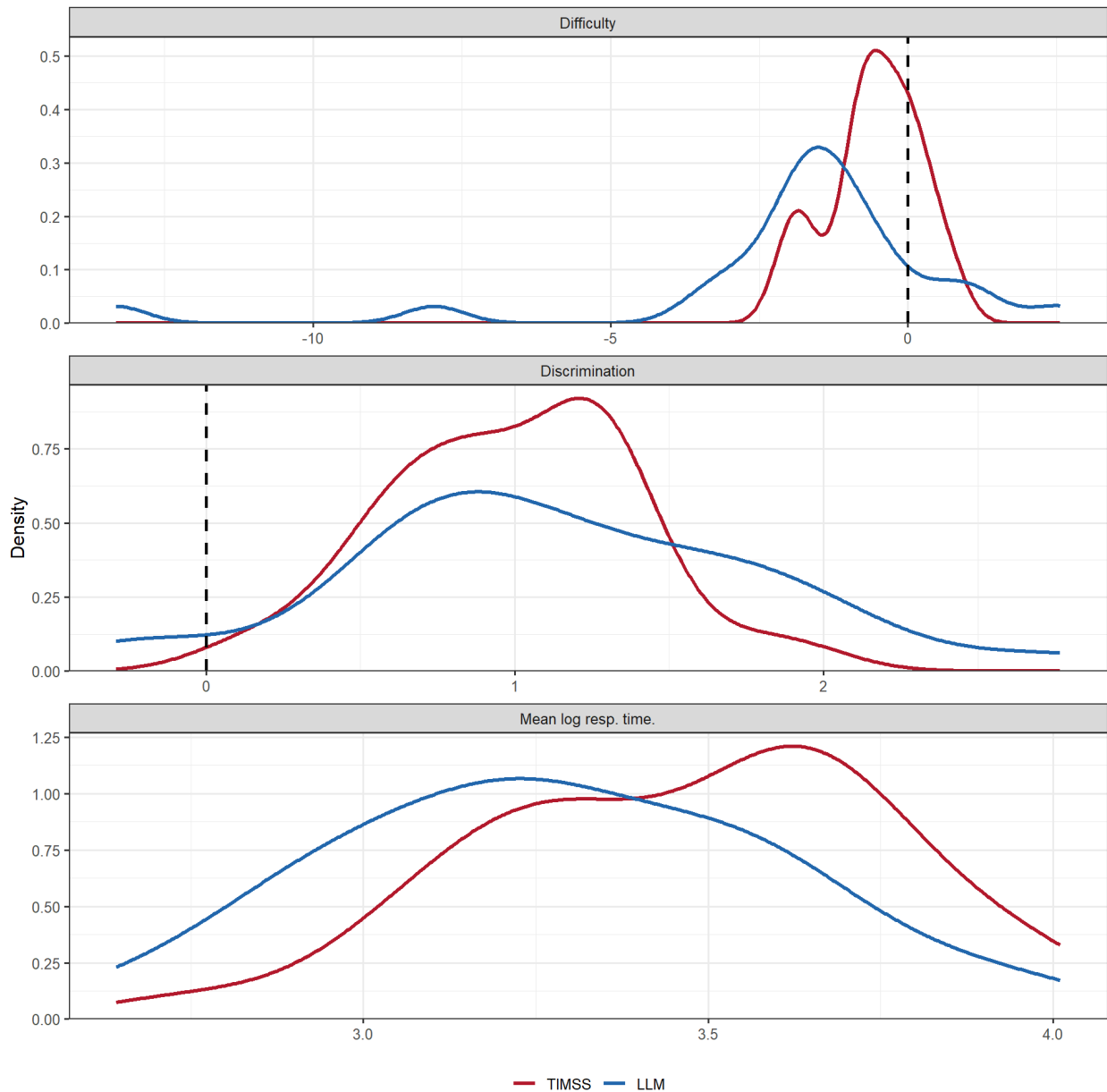
Distribution of Item Parameters for Mathematics: Constructed-Response Items by Generation Setup



This comparison leads to similar conclusions as above. It is noteworthy that the LLM-complex setup did not lead to generating more difficult items as the other two LLM setups, instead, it led to a bimodal distribution of item difficulty parameters. All LLM setups are characterised by the same problem with item discrimination - too many items with poor discrimination parameter (< 0.5).

Figure 19

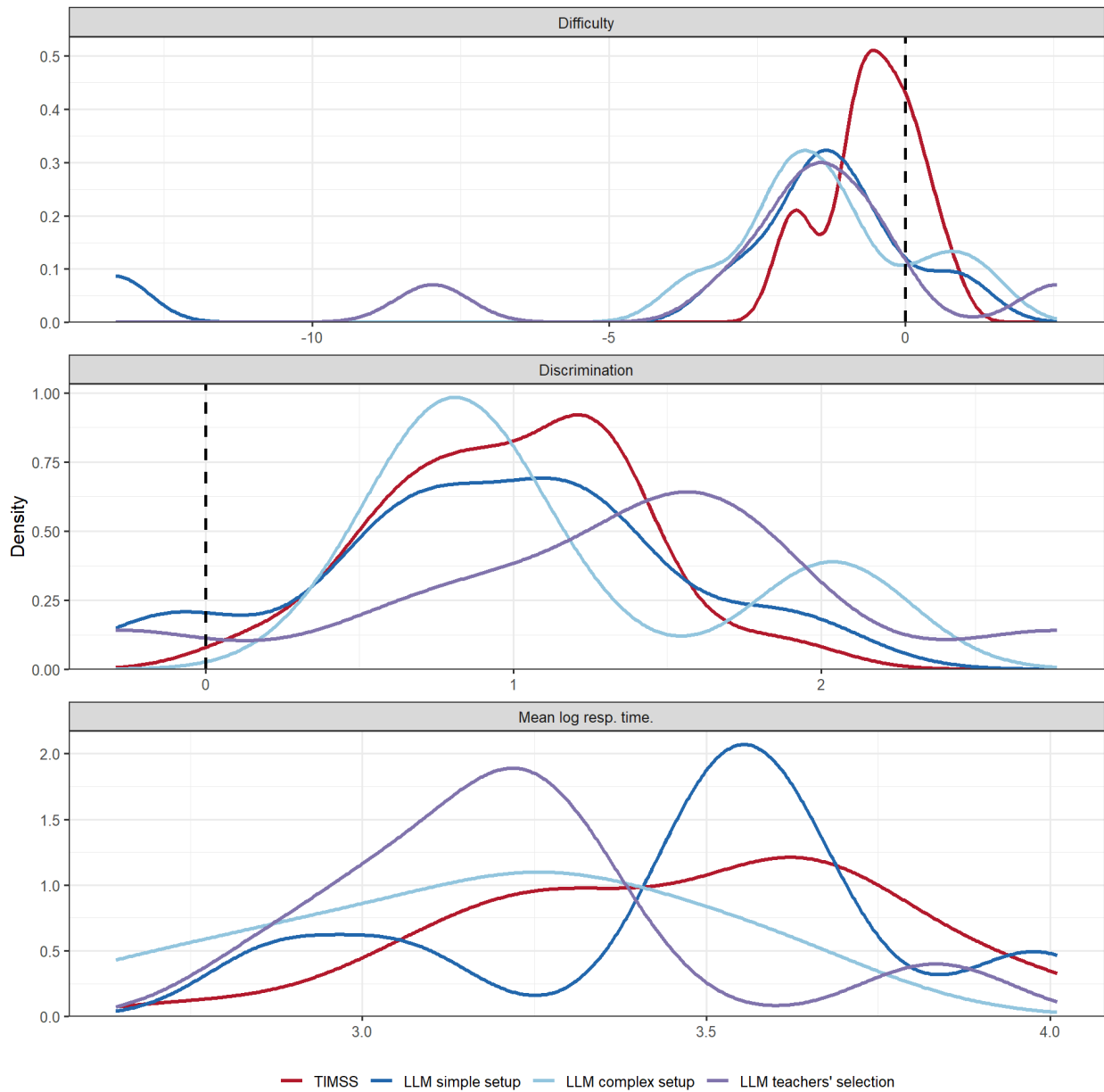
Distribution of Item Parameters for Mathematics: Multiple-Choice Items



LLM-generated MC items are characterised by lower difficulty, more platykurtic distribution of discrimination with fewer items in the intended range (0.5, 1.5) and more of very high and unusually high discrimination (above 1.5 and above 2.0), and faster mean logged response times.

Figure 20

Distribution of Item Parameters for Mathematics: Multiple-Choice Items by Generation Setup



The conclusions do not differ much when distributions are analysed comparing LLM-generation setups. Item difficulty parameters distribution is very similar among all LLM setups and again the LLM-complex setup failed to generate markedly more difficult items. LLM-teachers' selection items yielded lower mean logged response times and LLM-simple items higher mean logged response times than other item groups.

These differences were quantified again with the use of Kruskal-Wallis rank sum test that did not yield any statistically significant results in case of CR items. Exact test values are presented in Table S.4 in the Supplement. However, the analysis for MC items yielded differences on the verge of statistical significance for difficulty ($\chi^2(3) = 7.62, p = 0.055$) and mean logged response time ($\chi^2(3) = 8.53, p = 0.036$). In order to qualify these differences, we conducted a series of post-hoc comparisons, to which we used a Wilcoxon rank sum exact test. These tests confirmed that TIMSS MC mathematics items were more difficult than any other group of items, except the LLM-complex setup, where the difference was on the verge of significance. TIMSS Mc mathematics items were also characterised by longer response times. The Wilcoxon tests did not reveal statistically significant differences between the LLM setups in difficulty or mean logged response times. A detailed presentation of these analyses is in Tables S.5 and S.6 in the Supplement.

In the case of science the LLM-generated CR items proved to be much more difficult than the TIMSS items with a lot more exceedingly difficult items (difficulty > 2.0). LLM-generated items noted more items with high discrimination and lower mean logged response times than TIMSS items, the latter probably a consequence of differences in item difficulty. These distributions are presented in Figures 21 and 22.

Figure 21

Distribution of Item Parameters for Science: Constructed-Response Items

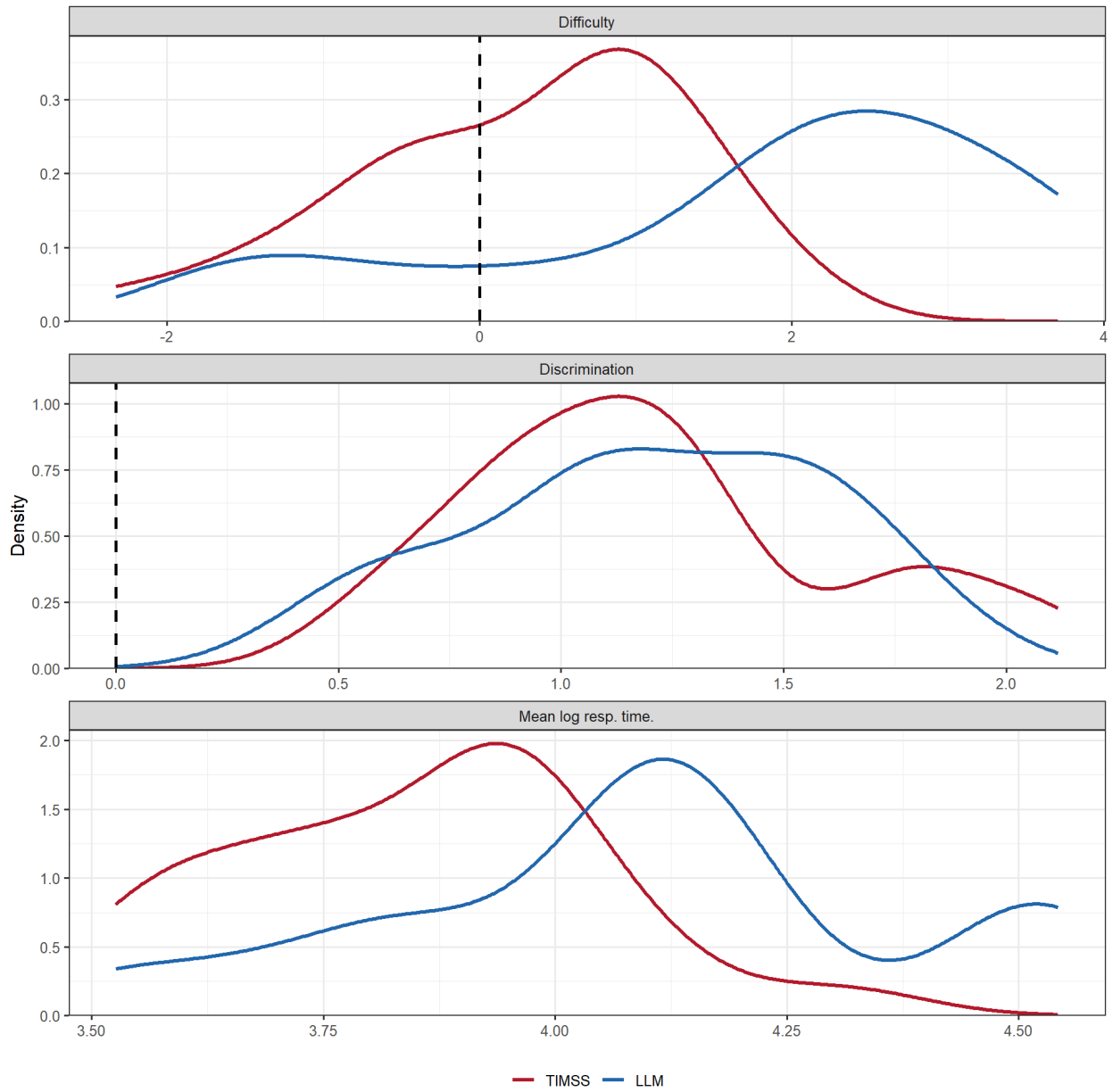
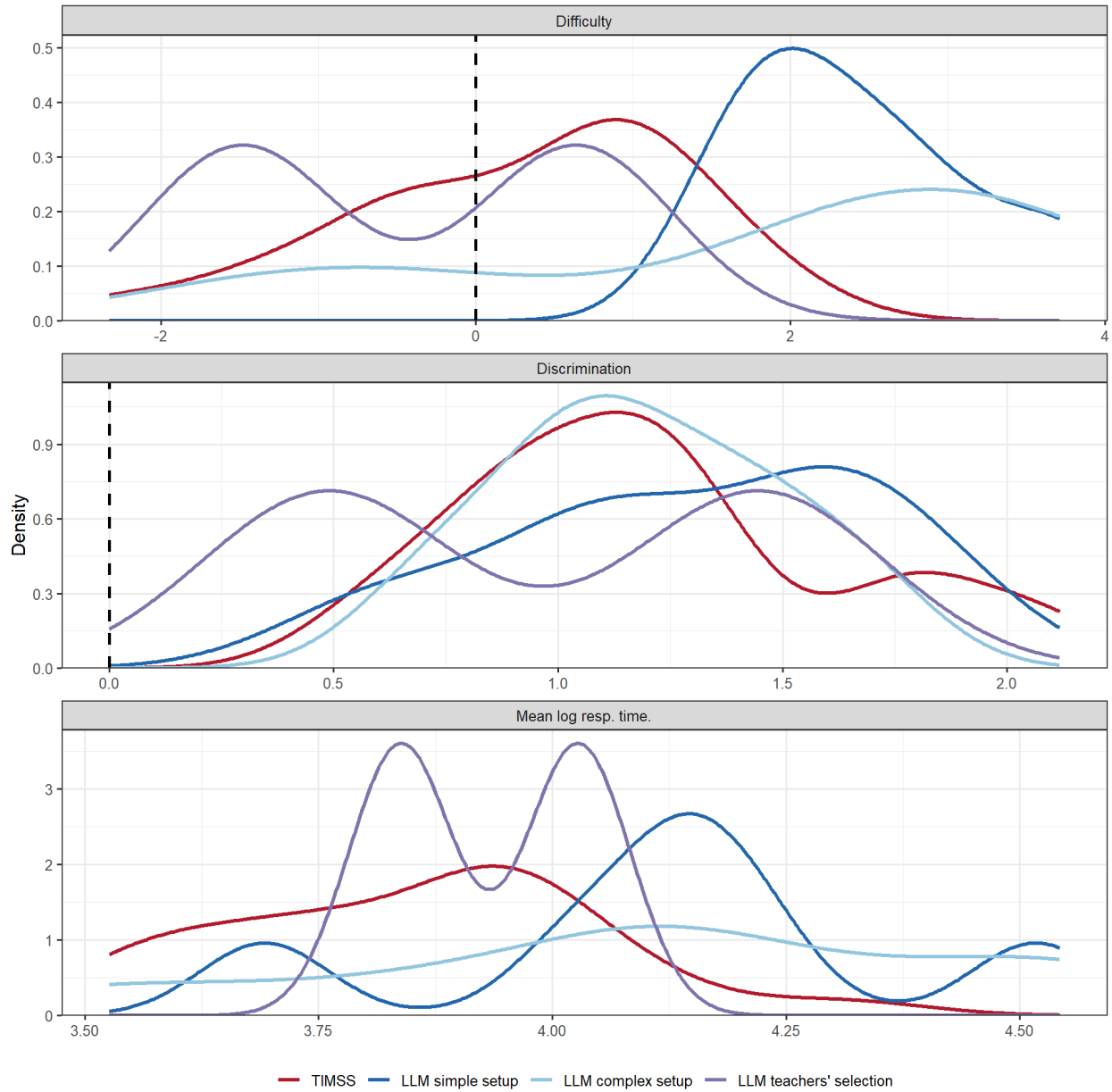


Figure 22

Distribution of Item Parameters for Science: Constructed-Response Items by Generation Setup



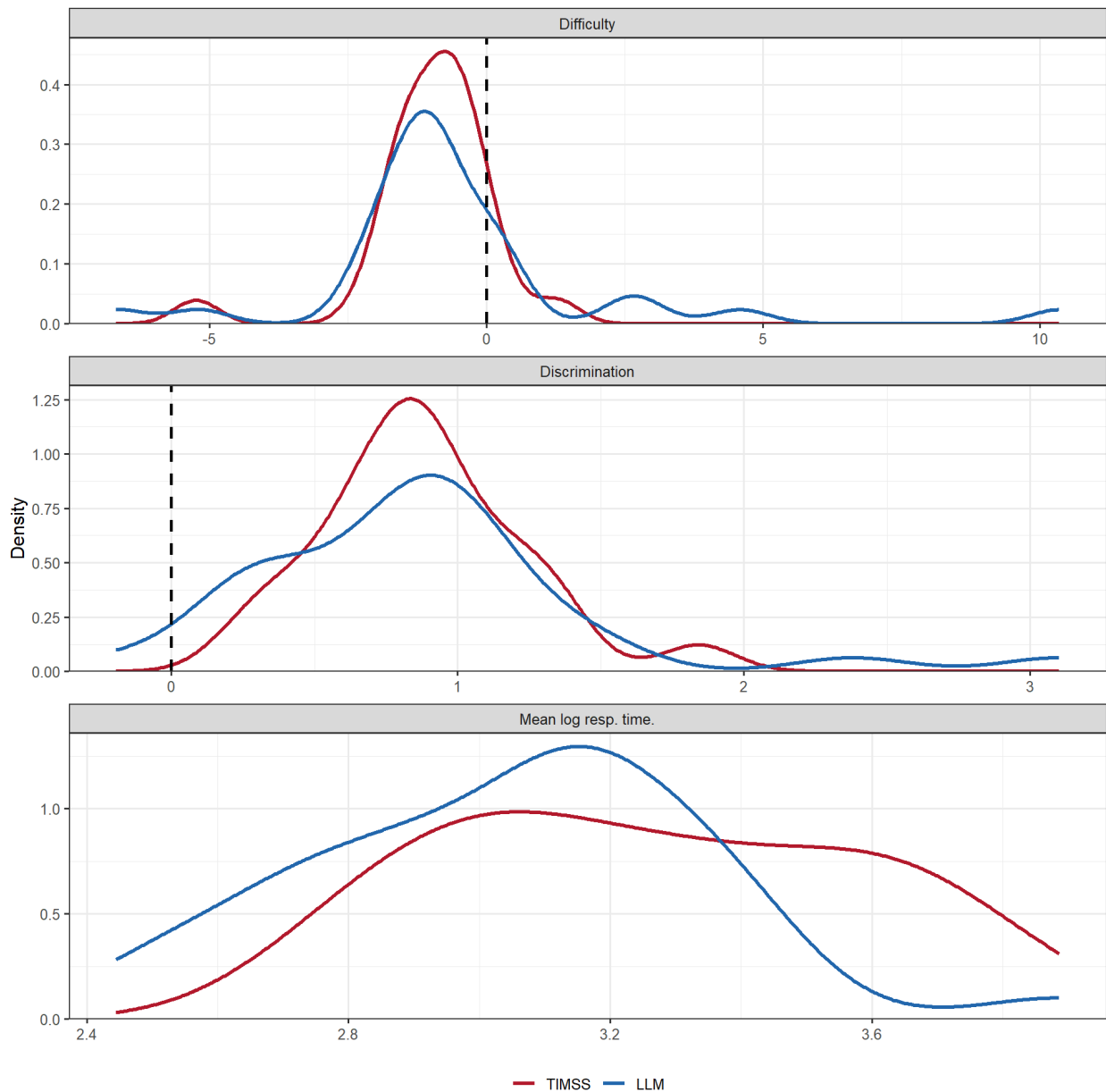
Unexpectedly, the LLM-simple setup produced very difficult items, at least in the case of CR science items. In contrast, the other two LLM setups yielded items of lower difficulty than TIMSS items. This is a bit surprising in case of the LLM-complex setup, which was supposed to generate “more challenging” items. It seems the setup did

succeed in generating a number of very difficult items (> 2), but also a number of easy and very easy items. All item groups yielded fairly similar discrimination parameter distributions, with a notable surplus of low discriminating items in the case of LLM-teachers' selection items. Other LLM setups did not seem to have these problems in the case of CR science items. The distribution of mean logged item response times revealed a similar pattern as the difficulty distribution - longest response times for the LLM-simple items and a bimodal distribution in case of the LLM-teachers' selection items, resembling its bimodal distribution of item difficulty parameters.

Results for MC science items are presented in Figures 23 and 24.

Figure 23

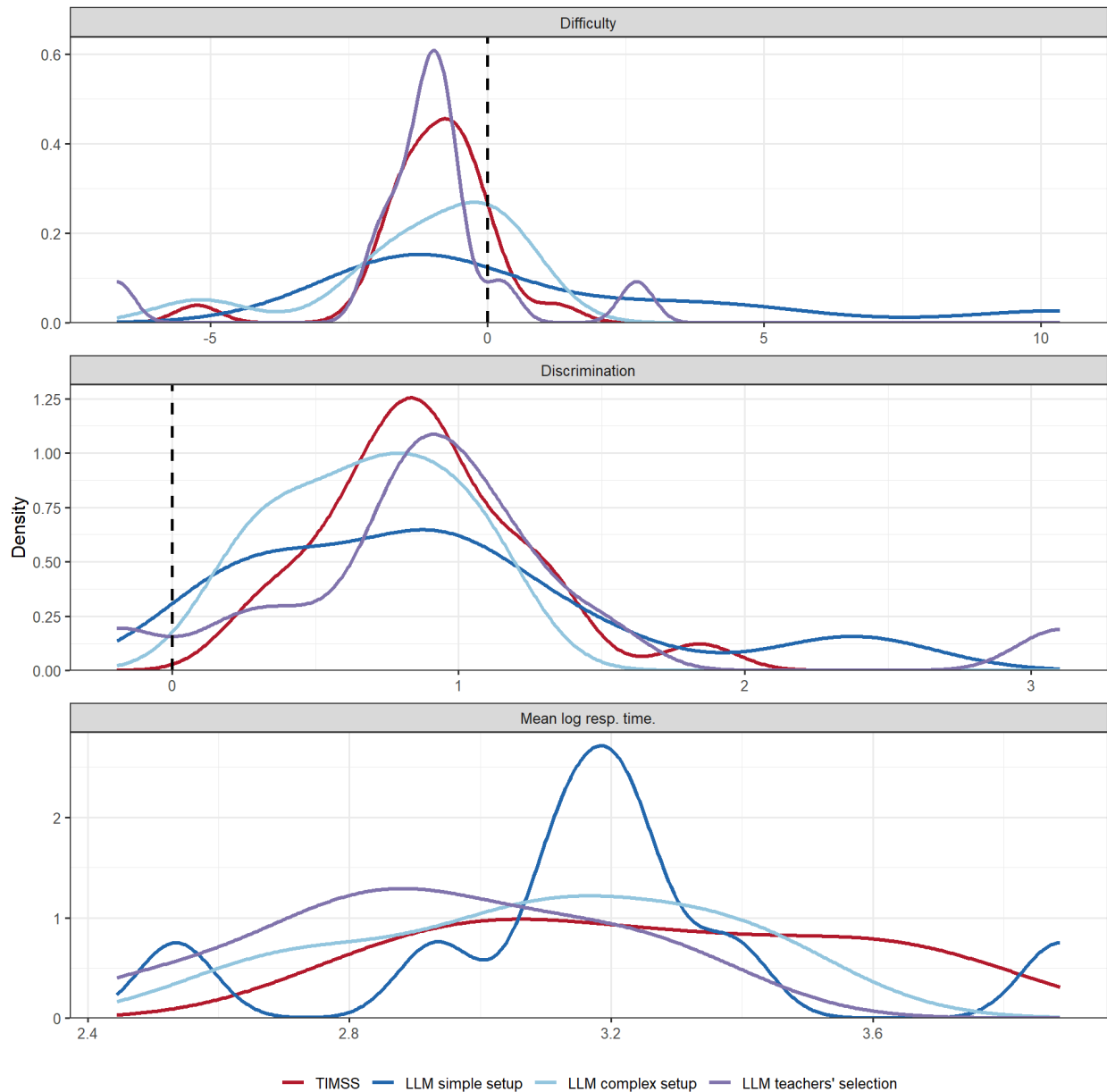
Distribution of Item Parameters for Science: Multiple-Choice Items



LLM-generated and TIMSS MC science items seem to follow the same distribution in the case of difficulty and discrimination parameters, with a notable surplus of very difficult items (> 2) and low discriminating items (< 0.5) in the case of LLM-generated items. In general, mean response time for LLM items is higher than for TIMSS items, but the pattern is reversed in the case of longer response times, when TIMSS items took on average longer to answer than LLM-generated questions.

Figure 24

Distribution of Item Parameters for Science: Multiple-Choice Items by Generation Setup



In the case of MC science items, the LLM-teachers' selection items followed a similar distribution of item difficulty parameters as TIMSS items, with a domination of fairly easy items (in the range of approximately -0.5, 0). The distribution for the LLM-complex setup was also similar, but with a more diverse difficulty of the items. The distribution for

the LLM-simple setup was very platykurtic, with both extremely easy and extremely difficult items. All item groups yielded similar discrimination parameters distributions, with the LLM-complex setup generating more items with low discriminations.

These differences were quantified again with the use of Kruskal-Wallis rank sum test that yielded significant results in the case of difficulty ($\chi^2(3) = 15.44, p = 0.001$) and response time ($\chi^2(3) = 8.61, p = 0.035$) for CR science items and in the case of response time for MC science items ($\chi^2(3) = 8.30, p = 0.040$). These results are presented in detail in Tables S.7 and S.8 in the Supplement.

In order to qualify these differences, we conducted a series of post-hoc comparisons, to which we used a Wilcoxon rank sum exact test. These tests confirmed that TIMSS CR science items were easier than all groups of LLM-generated CR science items. Moreover, TIMSS CR science items had lower mean response times than all LLM setups. A detailed presentation of these analyses is in Table S.9. Wilcoxon post-hoc tests also pointed to statistically significant differences in mean response times between TIMSS and LLM-generated items, with TIMSS items noting on average longer response times. These tests are presented in detail in Table S.10.

4.4.4 Differential Item Functioning

Differential Item Functioning (DIF) is one of the fundamental analyses in educational measurement, conducted to identify whether items function differently across various groups, even when individuals from these groups have equivalent levels of the underlying trait being measured. Testing items for DIF is important to maintain test fairness and validity.

Gender-related DIF is one of the most commonly assessed DIF types and we have also decided to perform this analysis. students' were classified to male and female genders on the basis of their responses to the socio-demographic questionnaire presented at the end of assessment.

The logistic regression method (Swaminathan and Rogers, 1990) was used to detect both uniform and nonuniform gender DIF. Because of the missing by design property of the data the EAP ability estimates (from unidimensional models estimated

separately for math and science, which included both TIMSS and LLM-generated item) were used as the matching criterion. Items with partial credit scoring rules were recorded into binary scores prior to the analysis by merging the middle score to the least frequently occurring of the remaining scores.

Because of the large number of performed tests, Holm's correction for multiple comparisons was applied to correct the p -values. The analysis was performed separately for maths and science items. We have used Jodoin and Gierl's (2001) criteria to categorise DIF effect size for the items for which DIF was statistically significant at the 0.05 significance level. Direction of DIF effect was reported on the basis of logistic regression coefficients. Analysis was performed using the *diffR* R package (Magis et al., 2010).

Only four items exceeded the threshold of Holm's corrected p -value of 0.05 and yielded DIF, all of them were classified as “moderate” DIF size on the basis of the Jodoin and Gierl's criteria (2001). Table 20 summarises the key information for this analysis, while Figure 25 displays the inter-group item characteristics curves. All of the DIF-flagged items were science items, two came from the original TIMSS group and two from the LLM-generated group, one from LLM-complex and from LLM-teachers' selection setup. Three of the flagged items favoured males (more difficult for females), one favoured females (more difficult for males). All of the items that favoured males involved electrical conductivity, while the item that favoured females related to maintaining good health. All of the flagged items had lower discrimination in the female group. Full results of DIF analysis for all items are presented in Tables S.11 and S.12.

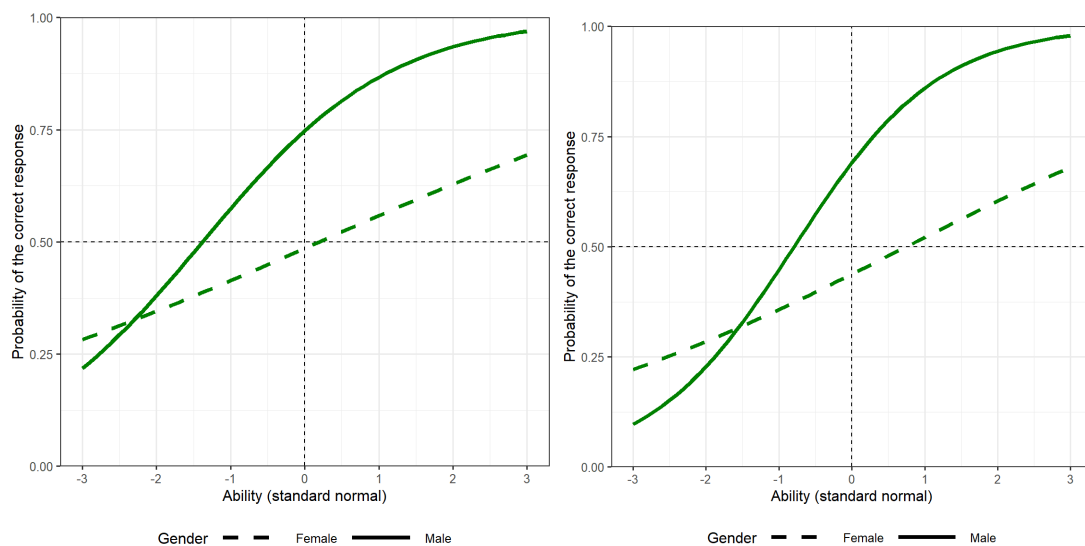
Table 20
Summary of DIF-Flagged Items

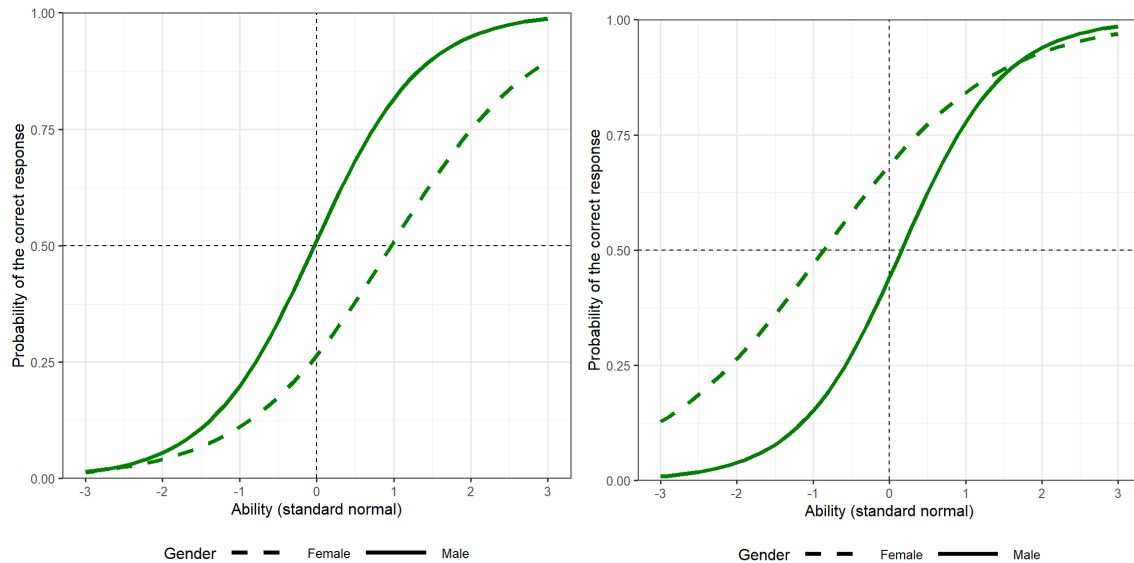
Source	Item	Statistic	P-value	Holm's corrected p-value	Δ Nagelkerke's R ²	Effect size (Jodoin & Gierl)	More difficult for	Higher discrimination for
LLM	S890009	19.69	0.000	0.005	0.058	moderate	males	males

Source	Item	Statistic	P-value	Holm's corrected p-value	Δ Nagelkerke's R ²	Effect size (Jodoin & Gierl)	More difficult for	Higher discrimination for
LLM	S09005 1	17.99	0.000	0.012	0.051	moderate	females	males
TIMSS	S05112 1	17.64	0.000	0.014	0.051	moderate	females	males
TIMSS	S06103 1	17.06	0.000	0.018	0.049	moderate	females	males

Figure 25

Graphical Display of DIF-Flagged Item Characteristics





Note. Number of the displayed items, from the left upper panel: S090051 (LLM), S051121 (TIMSS), S061031 (TIMSS), S890009 (LLM).

4.4.5 Testing Local Item Independence

The principle of local item independence is an important element of the standard latent variable models, including IRT models. It states that responses to different items should be statistically independent (not correlated), after accounting for an individual's latent trait level (for this study, their proficiency in mathematics or science). In other words, given a person's ability, the response to one item should not influence or predict the response to another. This ensures that the latent trait fully explains the relationships among item responses (see Vermunt & Magidson, 2004). Violations of the assumption can have adverse implications, most notably, it distorts person ability estimates' standard errors and can bias item parameters estimates, typically inflating discrimination parameters (and thus, any related to them, reliability/measurement precision indices). It can also signal that the test is not unidimensional and at least some of the items are influenced by an additional latent trait(s).

Local dependencies were assessed using the G^2 statistic (Chen & Thissen, 1997) on the basis of the same unidimensional math and science two-parameter IRT models involving both TIMSS and LLM-generated items, which parameters were described

above. Please note that because of the missing-by-design property of the data, local dependency statistics can be computed only for the pairs of items that were placed in the same booklet (i.e. there are students who responded to both items in a given pair). Consequently, local dependency statistics were computed for 1633 out of the total 4465 pairs of the calibrated math items and for 1645 out of the total 4465 pairs of the calibrated science items. Computations were performed using the residuals() function from the *mirt* R package (Chalmers, 2012).

Holm’s correction for multiple comparisons was applied separately to the results of math item pairs local dependency tests and to the results of science item pairs local dependency tests. After the correction local dependencies significant at the 0.05 level are presented in Table 21. Only two pairs of items breaking the local independence principle were identified, two from mathematics and two from science items. All of the identified items were LLM-generated items.

The breaking of local independence can be explained by analysing the items’ content. In the case of the first pair of mathematics items both items entailed calculating the area of a rectangle with sides of 8 cm and 5 cm. In the second pair both items also dealt with calculating the area of a rectangle. In the case of science, all flagged items measured knowledge about electrical conductance and all were MC items.

Table 21
Items Breaking Local Independence Principle

Item 1	Item 2	Correlation	G ²	df	p-value
Math					
M100027 (LLM, CR)	M890030 (LLM, MC)	0.289	26.64	1	0.000
M890009 (LLM, CR)	M100024 (LLM, CR)	0.262	21.92	2	0.028

Item 1	Item 2	Correlation	G ²	df	p-value
Science					
S090080 (LLM, MC)	S890038 (LLM, MC)	0.309	26.69	1	0.000
S890022 (LLM, MC)	S890038 (LLM, MC)	0.290	23.49	1	0.002

Note. Correlation – correlation coefficient between standardised residuals, *p*-value – Holm’s-adjusted *p*-value.

4.4.6 Summary of psychometric analyses

In order to summarise the key results of the psychometric analyses, we have decided to formulate a set of criteria, on which we could further assess the psychometric quality of the LLM-generated items. The criteria to flag an item were:

- *Negative discrimination*: $a \leq 0$
- *Low discrimination*: $a < 0.5$
- *Too easy*: $b < -2$
- *Too difficult*: $b > 2^6$
- *Gender-DIF*: gender-DIF detected (statistically significant at the 0.05 level in logistic regression DIF analysis)
- *Poor item fit*: item RMSE statistically significantly different from 0 at the 0.05 level
- *Local item dependence*: local dependency with some other item detected
- *Non-functional distractors*: this criterion applied only to MC items with easiness < 85%: at least one distractor chosen by less than 5% of respondents (Attali & Fraenkel, 2000).

⁶ We have used these thresholds as typical for the difficulty range of items in TIMSS, e.g. TIMSS 2011 featured only 5 items outside this range, TIMSS 2015 2 items, and TIMSS 2019 none such items. However, the trend to include a larger part of easy items changed that proportions for TIMSS 2023 were 20 items were identified outside the range. It is worthy to note that most of these items exceeded the threshold only very slightly (e.g. 2.1).

In the case of maths items, eighteen TIMSS items displayed at least one problem, in comparison to twenty-three LLM-generated items. The main problem identified in the TIMSS mathematics items was non-functional distractors (distractors chosen by less than 5% of participants). None of these items yielded a negative discrimination, but three items had a low discrimination parameter (< 0.5), in comparison, as much as three LLM-generated items displayed negative discrimination and two more low discrimination. The TIMSS items were also better fitting with regard to difficulty, with one too easy and one too difficult item, while this problem was much more pronounced among LLM items, with as much as nine too easy items (mainly MC) and five too difficult (mainly CR). Poor item fit was also more pronounced among the LLM-generated items, with seven flagged items, in contrast to only two TIMSS items. Among the LLM setups, the LLM-teachers' selection exhibited more issues than the other, especially the LLM-complex setup seemed to generate items with fewer serious problems (e.g. negative discrimination, poor item fit), however, this setup displayed some problems with misadjusted item difficulty.

In the case of science items, TIMSS items exhibited low occurrences of problematic metrics, with 3 items identified as having low discrimination and 3 classified as too easy. Notably, 2 items showed gender-related DIF, and 3 had poor item fit. Non-functional distractors (9) were more prevalent, contributing to a total of 19 flagged issues across 48 items. Items generated through LLMs exhibited more challenges compared to TIMSS, with a total of 32 flagged items across a total of 48 items. Key issues included one item with negative and nine further items with low discrimination, five items classified as too easy, and thirteen items as too difficult. Gender-DIF and non-functional distractors were also notable, with two and nine flagged instances, respectively. Poor item fit was a similarly serious issue among TIMSS (3 cases) and LLM (4 instances) items.

In summary, LLM-generated items exhibited more problems than TIMSS items, with having the right level of difficulty being the most challenging issue. LLM mathematics items contained a large portion of too easy items, especially among MC items, while LLM science items often displayed an excessive level of difficulty, especially among CR items. LLM MC science items also displayed a large number of low discrimination parameters, while in other item categories, TIMSS and LLM were on a par in this criterion. Interestingly, TIMSS items displayed some non-negligible problems with non-functional

distractors, which were very similar in scale as in the case of LLM-generated items. There were not many differences in the item psychometric quality across the LLM setups, with a notable exception of LLM-teachers' selection science items, which were characterised by the most suitable item difficulty levels. It seems that SMEs were able to validly assess items' difficulty levels.

Tables 22 and 23 present detailed results of this summary for mathematics and science items, respectively. Detailed item parameters information is presented in Annex 1, which is available online. Annexes 2-4 present wording of different sets of LLM-generated items (in Polish and machine translated to English).

Table 22

Summary of Psychometric Problems for Maths Items

Category	N	Discrimination		Easiness		Gender DIF	Bad fit	Local dep.	Non-func. distract.	Total
		Neg.	Low	Too easy	Too diffic.					
TIMSS										

Category	N	Discrimination		Easiness		Gender DIF	Bad fit	Local dep.	Non-func. distract.	Total
		Neg.	Low	Too easy	Too diffic.					
Overall	47	0	3	1	1	0	2	0	14	18
CR	14	0	0	0	1	0	1	0	0	2
MC	33	0	3	1	0	0	1	0	14	16
LLM – all setups										
Overall	48	3	2	9	5	0	7	0	7	23
CR	23	1	2	2	4	0	3	0	0	8
MC	25	2	0	7	1	0	4	0	7	15
LLM Simple Setup										
Overall	16	1	0	2	1	0	2	0	3	6
CR	8	0	0	0	1	0	0	0	0	1
MC	8	1	0	2	0	0	2	0	3	5
LLM Complex Setup										
Overall	16	0	1	3	2	0	1	0	2	7
CR	8	0	1	1	2	0	0	0	0	3
MC	8	0	0	2	0	0	1	0	2	4
LLM Teachers' Selection										
Overall	16	2	1	4	2	0	4	0	2	10
CR	7	1	1	1	1	0	3	0	0	4
MC	9	1	0	3	1	0	1	0	2	6

Table 23

Summary of Psychometric Problems for Science Items

Category	N	Discrimination		Easiness		Gender DIF	Bad fit	Local dep.	Non-func. distract.	Total
		Neg.	Low	Too easy	Too diffic.					
TIMSS										
Overall I	48	0	3	3	0	2	3	0	9	19
CR	23	0	0	1	0	1	2	0	0	4
MC	25	0	3	2	0	1	1	0	9	15
LLM – all setups										
Overall I	48	1	9	5	13	2	4	0	9	32
CR	16	0	1	0	9	1	1	0	0	12
MC	32	1	8	5	4	1	3	0	9	20
LLM Simple Setup										
Overall I	16	0	3	1	7	0	2	0	3	12
CR	7	0	0	0	4	0	1	0	0	5
MC	9	0	3	1	3	0	1	0	3	7
LLM Complex Setup										
Overall I	16	0	3	2	5	1	0	0	3	12
CR	7	0	0	0	5	1	0	0	0	6
MC	9	0	3	2	0	0	0	0	3	6
LLM Teachers' Selection										
Overall I	16	1	3	2	1	1	2	0	3	8
CR	2	0	1	0	0	0	0	0	0	1
MC	14	1	2	2	1	1	2	0	3	7

5. General discussion

5.1 Results summary and conclusion

Generating mathematics and science items that followed the TIMSS assessment framework with the use of LLMs proved to be a challenging task. The first encountered problem was high repetitiveness – the model created many the same or highly similar items, often replicating similar variations of the same item measuring a given content, e.g. electrical conductance. It also tended to revolve around the same or very similar numbers in mathematics items. Around 10% of the generated items were close duplicates and they mainly came from the zero- and one-shot setups.

The face validity of the generated items was not great, either. Some of the generated items were very simple and lacked the creativity, flair, and intriguing form of many of the original TIMSS items. Once in a while we have also spotted outright absurd or awkward items that did not bring much promise of valid use in an educational assessment of any kind (e.g.: *Sophie noticed that her dog always eats all the fruit that has fallen from the tree in her garden. One day she noticed that the dog had not eaten some fruit that was lying near the tree trunk. Give two reasons why the dog might not have eaten the fruit.* Another example: *We know that the number 84 is divisible by 4. What is the largest number that can be added to 84 so that the result is still divisible by 4?* This item is probably unsolvable without resorting to “higher” mathematics).

These observations were corroborated by SMEs ratings, where LLM-generated items were rated as less attractive and of less adjusted difficulty in comparison to real TIMSS items. In their comments, experts pointed out that many of the generated items were overly simplistic, very easy to solve for fourth-graders, as they were tapping into content covered in grades 1-3. On the other hand, many of the items contained terms or concepts that were well beyond any programme for grade 4, making the items too difficult for TIMSS assessment. Many of the generated items also displayed a wide variety of problems with the response options, e.g. the same distractor was repeated, there were two or more correct answers, or the correct answer was missing altogether. In many cases the correct option (key) was very obvious and stuck out from other response options, sometimes the options were not justified well and did not correspond to the common errors or misconceptions among grade four students.

Psychometric analysis showed that the most important challenge for the LLM-generated items was fitting into the required difficulty level. A challenge many of them failed. Noticeably, items from two domains suffered from different problems – mathematics items, especially MC, were too easy, while science items, especially CR, were too difficult. In general, both according to SMEs ratings and psychometric analyses, the generated science items displayed lower quality than mathematics items, corroborated by, e.g. low discrimination parameters.

We have also aimed to test various item generation setups. In the item generation process for the SMEs panel, we have tested setups differing on the number of provided item examples (real TIMSS items were used to this end). The setups featured zero, one, three or five examples. The results of the SMEs panel were not conclusive, although it seems that the three-shot setup generated the best answers.

As many items were unfavourably rated by the SMEs, we have decided to generate a new set of items using the “best” three-shot setup, but in two different variations: simple and complex. In the “complex” variant we have prompted the model to generate “more challenging” items, and designed the prompt to generate either multiple-choice or open-ended items (which was up to the LLM to decide in the former case). Judging from the psychometric properties of the generated items, the complex setup did not systematically generate more difficult items than the other two generational setups.

An interesting sub-project was using GPT-4 to generate images and tables that accompany the items. In TIMSS grade 4 images and tables are not overly complicated but sometimes do perform an important role as they convey additional information, useful in solving the item. Based on the SMEs ratings, multimodal LLMs are able to generate correct and handy item images and other graphic elements, but this task is not easy and far from automatic. Mathematics or science, in contrast to e.g. reading comprehension or foreign language items, often require high-quality images.

Due to the instability of the DALLE-3 mode it was hard to specify the details of the images such as the precise number of objects in the image or their exact shapes, as DALL-E 3 would frequently reinterpret the prompt in stylistic ways that broke those strict constraints. This sparked the need to use the embedded python execution capabilities of the GPT4 model, which with the use of Matplotlib - a mature Python library that

programmatically renders vector graphics and charts, was able to generate the images that required more fine-grained control of their content. Due to the need of relying on the team members' intuition of when to use one mode of image generation over the other and the repetitive process of refining and correcting the content of required images this task proved laborious and hardly possible to fully automate. However, without access to DALLE-3 and the Matplotlib capabilities embedded with the GPT4 model, this task would have proven far more laborious. Therefore, it can be safely said that using AI tools for this task made it significantly easier to accomplish.

Of course, our analyses present a “worst-case scenario”, as the generated items did not undergo any human editing⁷ before going for SMEs or psychometric assessments. We also did not pick from the generated items but instead present for assessments a previously selected number of items, discarding only outright incorrect or repetitive items (see supplement and Annex 4 for more information on this topic). Hence, our study was more similar to a pilot administration, serving to validate the items at hand and select the best performing ones for the final item pool, than to an “item quality” pageant. Moreover, the generated items were compared to real TIMSS tasks – the world's aristocracy of items, developed, edited, validated, and polished in a number of studies by leading researchers and experts.

We also did not aim to develop a fully automatic method for developing high-quality items without any human intervention – for doing that, our project was too small. The main aim was rather to test how good a set of items we can have with the use of common models and simple prompt-engineering techniques, but without large human assistance. As it seems, many of the LLM-generated items did not land that far from the TIMSS (high) quality threshold.

An important part of this study was to show what are the main problems encountered in the case of automatically generated items so that future projects would have an additional piece of information to start from. It is also important to show problems in AIG and provide open access to detailed information on the process of item generation and its results. The field is not specifically rich in such problem-centred research as much

⁷ Save for the iterative generation of images.

of the AIG research is concentrated on commercial success, hence it is not exposing all of its nuts and bolts to the public (and it is not eager to present AIG problems either).

5.2 Future directions

The field of AIG is going to change dynamically in the forthcoming years and a similar dynamic development will certainly be enjoyed by LLMs. Capabilities of the models increase every month, and new possibilities are added to old models. What was impossible when we wrote the grant proposal (e.g. solving mathematical problems in LLMs, Drori et al., 2022), became a standard when we answered the reviews a few months afterwards (e.g. Testolin, 2024). The dynamic increase in LLMs capabilities will certainly enhance the possibility of AIG.

Many research questions have to be answered though, before AIG will be 100% successful in generating high-quality educational assessment items. For example, the question of optimal values for key model parameters, such as temperature, is still open and up to be tested empirically, as it was set to 0.8 based only on a short pilot study. All of the other hyperparameters were kept at their default values. A proper exploration of their influence would require a study of a much bigger scale than that currently presented. This is a great and valuable idea for future research.

Similarly, we do not know how far we can go with prompt engineering, and how we can increase the quality of output by resorting to other techniques supporting prompting such as iterative prompting, contrastive learning, providing additional context (cf. long context) or knowledge (e.g. through the framework of retrieval-augmented generation, RAG; Li et al., 2024). Fine-tuning and instruction-tuning are very likely to notably increase the AIG quality (Gorgun & Bulut, 2025). An interesting idea is yet another reincarnation of item cloning (Gierl & Lai, 2015), i.e. generating different item versions based on one template. With the use of LLMs we can now not only generate very close item copies, e.g. for the purpose of parallel tests, but also items that keep the idea of the exemplary item and develop it further creatively.

In our prompting scheme, we aimed to balance informativeness with stability, given empirical evidence that long prompts can dilute the signal. Accordingly, we adopted evidence-based guidelines for long-context use and kept prompts lean, with key elements

front-loaded and reiterated at the end, reinforced by clear structure (cf. Liu et al., 2023). Of course, more prompt-engineering research is also needed, as iterative prompting and other techniques can help to improve the quality of even very long prompts (Zhang et al., 2024).

Creating graphical elements for items is another challenge. In our study, we were able to generate large quantities of items in a blink of an eye, but the creation of images was slow, as we prompted the models on a trial-and-error basis, not always with great success, as often noted by the SMEs. It seems that recent integration of Python-based frameworks enabled plotting high-quality graphs and tables, but researching ways of efficient image creation is still needed.

Generating items of spot-on difficulty level is also a nontrivial task. Developing ways of providing models with better guidance in this regard seems one of the most needed challenges for the near future. Of course, this can vary in, *nomen est omen*, difficulty depending on the domain, the whole process of AIG seems a lot easier in language related domains and tasks, such as reading comprehension (Sayin & Gierl, 2024) or English as a second language (Attali et al., 2022), than in mathematics or science, without even mentioning specialised fields such as medical items (Bulut et al., 2024).

One promising avenue for future research is the development of more precise methods for controlling the difficulty of LLM-generated items. This could be achieved through tailored prompting strategies that incorporate example items of known difficulty levels. For instance, in few-shot prompting, one might include items with established IRT difficulty parameters within specific ranges (e.g., -2 to -0.5 for "easy" items, -0.5 to $+0.5$ for "average" items, and $+0.5$ to $+2$ for "difficult" items).

Greater control over the item generation process, for example, by separately generating items of different formats (e.g., multiple-choice and constructed-response), may also enhance the quality and consistency of the output. Simplifying and narrowing the scope of prompts could contribute to improved item quality, while also reducing the variability introduced by the inherent randomness of LLM-generated content.

Additionally, prompting models to verify key formal properties of the generated items—such as ensuring that only one correct answer is present and that all response

options are distinct—could further improve quality. Previous research has highlighted the frequent occurrence of issues related to response option plausibility in LLM-generated items (Kuusemets et al., 2024).

More advanced prompting strategies, such as Chain-of-Thought prompting, may also facilitate the generation of more cognitively demanding items, particularly in mathematics (Wei et al., 2023). Combining such prompting strategies with self-filtering and self-refinement prompts may further enhance the quality of the generated output (Hang et al., 2024). Initial evidence suggests that various machine learning techniques are already quite successful in predicting item difficulty and other characteristics (e.g. Štěpánek et al., 2023).

An alternative approach is fine-tuning, which involves training the model on a large set of high-quality item examples and a structured body of domain-specific knowledge. However, this method is resource-intensive and may be limited by the availability of suitable training data. A key advantage of fine-tuning is its potential to mitigate limitations in the model's domain knowledge. When item generation performance is unsatisfactory, fine-tuning can help disentangle whether the issue stems from suboptimal prompting or from the model's inadequate knowledge in the target domain. Despite all difficulties, we think that re-training the model with a large and diversified set of materials from different countries (e.g. teaching textbooks, didactical articles, presentations of curricula descriptions, learning objectives) along with a similarly large and diverse set of the example items is an excellent avenue for any future research.

If we were to recommend avenues for future research, we would suggest that the use of fine-tuned, in-domain Large Language Models (LLMs), especially those run locally (Bulut et al., 2024), is likely to yield the best results, particularly in specialised or low-resource domains such as science or medical assessments. Running models locally also addresses important concerns related to data privacy and output stability. In contrast, using off-the-shelf models offers no control over future changes or updates these models may undergo, meaning that even carefully developed and tested prompts may not produce consistent results over time.

A modular architecture of fine-tuned local models, with each component tailored to a specific stage of the test development pipeline, appears to be the most promising

solution, at least in the near term (which, in the rapidly evolving LLM landscape, may be relatively brief). This setup would allow for the automation of the entire test development process through a coordinated cascade of models, without compromising item quality. For instance:

- One (sub)model could generate items,
- Another could evaluate their quality (Gorgun & Bulut, 2025) and predict difficulty,
- A third could handle test assembly tasks, such as ensuring that similar items do not appear in the same booklet,
- A final model or agent could score student responses from pilot administrations, even for constructed-response or graphical formats (Tyack et al., 2024).

Together, such a system could form the foundation of a true "item factory" (von Davier et al., 2024).

However, despite this vision, it is likely to be a long time before we can significantly reduce the role of the human-in-the-loop. For now, human editing, selection, and review remain indispensable to ensure that automatic item generation (AIG) yields high-quality results. Without this intervention, most LLM-generated items are likely to remain in the "needs editing" zone, as illustrated by the image on the cover of this report.

LLMs remain, in many ways, black boxes, and their unpredictable behavior is well illustrated by one final example: how the model attempted to solve an absurd item it had generated itself.

We know that the number 84 is divisible by 4. What is the largest number that can be added to 84 so that the result is still divisible by 4?

Prompt: *Answer the question and then estimate to what level of education it is suitable.*

Answer: We know that the number 84 is divisible by 4. What is the largest number that can be added to 84 so that the result is still divisible by 4?

Given that 84 is divisible by 4, any number added to 84 must also be divisible by 4 for the sum to remain divisible by 4. The largest such number is 4 itself, as adding any multiple of 4 to 84 will keep the sum divisible by 4.

Explanation:

A number is divisible by 4 if the number formed by its last two digits is divisible by 4. Adding 4 to 84 results in 88, which ends in '88'. Since 88 is divisible by 4, the entire number is divisible by 4. Adding any number greater than 4 that is divisible by 4 (e.g., 8, 12, 16) would also result in a sum divisible by 4, but the problem specifies finding the largest number that can be added to 84, which is 4 itself.

Educational Level:

This problem involves understanding and applying the divisibility rule for 4, which is typically introduced around 5th or 6th grade. Therefore, this problem is suitable for students in upper elementary or early middle school, approximately 5th to 6th grade.

Item:

Sophie noticed that her dog always eats all the fruit that has fallen from the tree in her garden. One day she noticed that the dog had not eaten some fruit that was lying near the tree trunk.

Give two reasons why the dog might not have eaten the fruit.

Answer:

The fruit might be rotten or spoiled.

Dogs are less likely to eat fruit that is overripe, rotten, or has an unpleasant smell.

The fruit might not be safe or appealing to the dog.

Some fruits, especially those with hard or bitter skins, may not appeal to dogs. Additionally, the dog might have sensed that the fruit could be harmful or toxic.

Educational Level Suitability:

This task requires basic logical reasoning and understanding of animal behavior. It is suitable for grades 3-6 in elementary school.

6. References

- Attali, Y., & Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement*, 37(1), 77-86.
- Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5, 903077.
- Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5, 100161.
- Bulut, O., Gorgun, G., & Tan, B. (2024). Item Difficulty and Response Time Prediction with Large Language Models: An Empirical Analysis of USMLE Items. *Preprint*. <https://era.library.ualberta.ca/items/415e9fe5-c3e3-46c6-8725-258f9a63e617/download/dbfb6e3c-9319-446a-89a6-b59fa1220f98>
- Chae, Y. M., Park, S. G., & Park, I. (2019). The relationship between classical item characteristics and item response time on computer-based testing. *Korean journal of medical education*, 31(1), 1-9. <https://doi.org/10.3946/kjme.2019.113>
- 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>
- Chen, W. H. and Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.
- Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., ... & Strang, G. (2022). A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32), e2123433119.
- Embretson, S. (2021). Response Time Relationships Within Examinees: Implications for Item Response Time Models. In: Wiberg, M., Molenaar, D., González, J., Böckenholt, U., Kim, JS. (eds) *Quantitative Psychology. IMPS*

2020. *Springer Proceedings in Mathematics & Statistics*, vol 353. Springer, Cham. https://doi.org/10.1007/978-3-030-74772-5_5
- Falotico, R., and Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2), 463–470. <https://doi.org/10.1007/s11135-014-0003-1>
- Gierl, M. J., & Lai, H. (2015). Automatic item generation. In *Handbook of test development* (pp. 410-429). Routledge.
- Gorgun, G., & Bulut, O. (2025). Instruction-Tuned Large-Language Models for Quality Control in Automatic Item Generation: A Feasibility Study. *Educational Measurement: Issues and Practice*, 44(1), 96-107. <https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12663>
- Gwet (2014). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. 4th ed. Gaithersburg, MD: Advanced Analytics.
- Hang, C. N., Tan, C. W., & Yu, P. D. (2024). MCQGen: A large language model-driven MCQ generator for personalized learning. *IEEE Access*, 12, 102261-102273, [10.1109/ACCESS.2024.3420709](https://doi.org/10.1109/ACCESS.2024.3420709)
- Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., ... & Sharma, M. (2024). Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*.
- Jodoin, M. G. and Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education* 14(4), 329-349. https://doi.org/10.1207/S15324818AME1404_2
- Kay, C. S. (2024). Extraverted introverts, cautious risk-takers, and selfless narcissists: A demonstration of why you can't trust data collected on MTurk. <https://osf.io/zs6pk/download>
- Kuusemets, L., Parve, K., Ain, K., & Kraav, T. (2024). Assessing AI-generated (GPT-4) versus human created MCQs in mathematics education: A comparative inquiry into vector topics. *International Journal of Education in Mathematics, Science, and Technology (IJEMST)*, 12(6), 1538-1558. <https://doi.org/10.46328/ijemst.4440>

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.
<https://pdfs.semanticscholar.org/7e73/43a5608fff1c68c5259db0c77b9193f1546d.pdf>
- Li, X., Cao, Y., Ma, Y., & Sun, A. (2024). Long Context vs. RAG for LLMs: An Evaluation and Revisits. *arXiv preprint arXiv:2501.01880*, <https://arxiv.org/abs/2501.01880>
- Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, 10, 1533.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.
- Magis, D., Beland, S., Tuerlinckx, F. and De Boeck, P.. (2010). A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning. *Behavior Research Methods* 42, 847-862.
<https://doi.org/10.3758/BRM.42.3.847>
- Mullis, I.V.S., Martin, M.O., Cotter, K.E., and Centurino, V.A.S. (2017). TIMSS 2019 Item Writing Guidelines. Retrieved from Boston College, TIMSS & PIRLS International Study Center website:
<https://timssandpirls.bc.edu/timss2019/pdf/T19-item-writing-guidelines.pdf>
- Mullis, I.V.S., Martin, M.O., and von Davier, M. (Eds.). (2021). TIMSS 2023 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2023>
- Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models?. *arXiv preprint arXiv:2405.00492*.
- Pokropek, A. (2011). Missing by design: Planned missing-data designs in social science. *ASK. Research & Methods*, (20), 81-105.

- Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to Generate Reading Comprehension Items. *Educational Measurement: Issues and Practice*, 43(1), 5-18.
- Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19), 4104.
- Testolin, A. (2024). Can neural networks do arithmetic? a survey on the elementary numerical skills of state-of-the-art deep learning models. *Applied Sciences*, 14(2), 744.
- Tyack, L., Khorramdel, L., & von Davier, M. (2024). Using Convolutional Neural Networks to Automatically Score Eight TIMSS 2019 Graphical Response Items. *Computers and Education: Artificial Intelligence*, 100249.
- Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa—A comparison of basic properties. *MethodsX*, 10, 102212.
<https://doi.org/10.1016/j.mex.2023.102212>
- Verbic, S., & Tomic, B. (2009). Test item response time and the response likelihood. *arXiv preprint arXiv:0901.4356*. <https://doi.org/10.48550/arXiv.0901.4356>
- Vermunt, J. K., & Magidson, J. (2004). Local independence. In *The Sage encyclopedia of social sciences research methods* (pp. 580-581). Sage.
https://jeroenvermunt.nl/ermss2004c.pdf?utm_source=chatgpt.com
- von Davier, A. A., Runge, A., Park, Y., Attali, Y., Church, J., & LaFlair, G. T. (2024). The item factory: Intelligent automation in support of test development at scale. In H. Jiao & R. W. Lissitz (Eds.), *Machine learning, natural language processing, and psychometrics* (pp. 1–25). Information Age Publishing.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. <https://arxiv.org/abs/2201.11903>
- Willerman, B. (1955). The adaptation and use of Kendall's Coefficient of Concordance (W) to sociometric-type rankings. *Psychological Bulletin*, 52(2), 132–133. <https://doi.org/10.1037/h0041665>

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13, 1-7. <https://doi.org/10.1186/1471-2288-13-61>

Zhang, Z., Chen, R., Liu, S., Yao, Z., Ruwase, O., Chen, B., ... & Wang, Z. (2024). Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37, 60755-60775.

7. Supplement

7.1 Duplicates removed from the generated item set

Duplicates removed from the item set:

S090025

S090028

S090034

S090039

S090065

S090069

S090074

S090079

Items introduced to the item set in place of the removed duplicates:

S090080

Condition: Zero-shot

Subject: Science

Content Domain: Physical Science

Cognitive Domain: Knowing

Generated: Który z poniższych przedmiotów jest najlepszym przewodnikiem energii elektrycznej?

- A. Plastikowa łyżka
- B. Gumowa rękawica
- C. Srebrny widelec
- D. Szklana butelka

S090081

Condition: Zero-shot

Subject: Science

Content Domain: Life Science

Cognitive Domain: Applying

Generated: Które z poniższych zwierząt jest przykładem drapieżnika?

- A. Królik
- B. Krowa
- C. Wilk
- D. Ślimak

S090082

Condition: Zero-shot

Subject: Science

Content Domain: Physical Science

Cognitive Domain: Applying

Generated: Które z poniższych stwierdzeń najlepiej opisuje, co się dzieje z kostką lodu, gdy pozostawimy ją na stole w temperaturze pokojowej?

- A. Kostka lodu stanie się większa, ponieważ pochłonie ciepło z powietrza.
- B. Kostka lodu stopnieje, ponieważ pochłonie ciepło z powietrza.
- C. Kostka lodu stanie się cięższa, ponieważ pochłonie ciepło z powietrza.
- D. Kostka lodu stanie się lżejsza, ponieważ odda ciepło do powietrza.

S090083

Condition: Zero-shot

Subject: Science

Content Domain: Life Science

Cognitive Domain: Applying

Generated: W ogrodzie mieszka wiele zwierząt i roślin. Poniżej znajduje się tabela przedstawiająca kilka z nich oraz ich cechy.

Organizm	Rodzaj pokarmu	Miejsce zamieszkania
Pająk	Owady	Pajęczyna
Biedronka	Mszyce	Liście
Mysz polna	Nasiona, owady	Nory w ziemi
Sowa	Małe gryzonie, owady	Dziuple w drzewach
Dąb	Słońce, woda, minerały	Las

Zadanie: Wybierz dwa organizmy z tabeli, które mają ze sobą interakcje w ekosystemie. Opisz krótko, na czym polega ta interakcja.

(Przykład odpowiedzi: Pająk i biedronka - Pająk może zjadać biedronkę, ponieważ pająki jedzą owady.)

S090084

Condition: One-shot

Subject: Science

Content Domain: Life Science

Cognitive Domain: Applying

Generated: Która z poniższych cech nie jest typowa dla ssaków?

- A. Mają sierść lub futro
- B. Oddychają przez płuca
- C. Składają jajka
- D. Karmią młode mlekiem

S090085

Condition: One-shot

Subject: Science

Content Domain: Physical Science

Cognitive Domain: Knowing

Generated: Które z poniższych urządzeń wykorzystuje energię magnetyczną?

- A. Lodówka

- B. Radio
- C. Pralka
- D. Kompas

S090086

Condition: One-shot

Subject: Science

Content Domain: Life Science

Cognitive Domain: Knowing

Generated: Które z poniższych zdań najlepiej opisuje, co jest potrzebne roślinom do fotosyntezy?

{'A': 'Tlen i woda', 'B': 'Dwutlenek węgla i tlen', 'C': 'Dwutlenek węgla i światło słoneczne', 'D': 'Tlen i światło słoneczne'}

S090087

Condition: One-shot

Subject: Science

Content Domain: Earth Science

Cognitive Domain: Knowing

Generated: Przyjrzyj się następującemu rysunkowi, przedstawiającemu cykl życia wody.

[OBRAZEK: Cykl życia wody pokazujący parowanie, kondensację i opady deszczu]

Jak nazywa się proces, w którym woda z oceanów, jezior i rzek unosi się w powietrze?

- A) Kondensacja
- B) Parowanie
- C) Opady
- D) Krystalizacja

S090088

Condition: Five-shot

Subject: Science

Content Domain: Life Science

Cognitive Domain: Applying

Generated: Wypisz dwa sposoby, w jakie rośliny przystosowują się do życia w środowisku suchym.

- 1.
- 2.

7.2 Kruskal-Wallis Test Results

Table S.1

Kruskal-Wallis Test Results for Mathematics Items' Item Parameters - All Items

Parameter	Kruskal-Wallis rank sum	
	test	ANOVA
Discrimination	$\chi^2(3) = 0.35, p = 0.951$	$F(3, 91) = 0.11, p = 0.952$
Difficulty	$\chi^2(3) = 0.72, p = 0.867$	$F(3, 91) = 1.05, p = 0.373$
Mean log resp. time	$\chi^2(3) = 2.53, p = 0.471$	$F(3, 91) = 0.65, p = 0.588$

Note. This analysis compared TIMSS (N=47) vs. LLM-generated items (N=48).

Table S.2

Kruskal-Wallis Test Results for Science Items' Item Parameters - All Items

Parameter	Kruskal-Wallis rank sum	
	test	ANOVA
Discrimination	$\chi^2(3) = 1.05, p = 0.789$	$F(3, 91) = 0.35, p = 0.788$
Difficulty	$\chi^2(3) = 9.21, p = 0.027$	$F(3, 91) = 5.61, p = 0.001$
Mean log resp. time	$\chi^2(3) = 12.41, p = 0.006$	$F(3, 91) = 4.78, p = 0.004$

Note. This analysis compared TIMSS (N=48) vs. LLM-generated items (N=47).

Table S.3

Wilcoxon Post-Hoc Test Results for Science Items' Parameters - All Items

Comparison	Wilcoxon rank sum exact	
	test	t-test
Difficulty		
TIMSS vs. LLM (all items)	$W = 1055.00, p = 0.591$	$t(65.1) = -1.35, p = 0.181$
TIMSS vs. LLM simple setup	$W = 217.00, p = 0.020$	$t(15.4) = -2.37, p = 0.031$
TIMSS vs. LLM complex setup	$W = 326.00, p = 0.376$	$t(18.1) = -0.90, p = 0.378$
TIMSS vs. LLM teachers' selection	$W = 512.00, p = 0.047$	$t(19.8) = 1.41, p = 0.173$

Comparison	Wilcoxon rank sum exact	
	test	t-test
LLM simple setup vs. LLM complex setup	$W = 142.00, p = 0.401$	$t(25.4) = 1.44, p = 0.162$
LLM simple setup vs. LLM teachers' selection	$W = 175.00, p = 0.030$	$t(22.2) = 2.86, p = 0.009$
LLM complex setup vs. LLM teachers' selection	$W = 173.00, p = 0.094$	$t(28.7) = 1.69, p = 0.102$
Mean log resp. time		
TIMSS vs. LLM (all items)	$W = 1352.00, p = 0.096$	$t(81.8) = 1.61, p = 0.111$
TIMSS vs. LLM simple setup	$W = 346.00, p = 0.829$	$t(18.5) = -0.08, p = 0.938$
TIMSS vs. LLM complex setup	$W = 398.00, p = 0.836$	$t(19.6) = 0.04, p = 0.970$
TIMSS vs. LLM teachers' selection	$W = 608.00, p = 0.000$	$t(24.5) = 4.04, p = 0.000$
LLM simple setup vs. LLM complex setup	$W = 125.00, p = 0.861$	$t(29) = 0.09, p = 0.930$
LLM simple setup vs. LLM teachers' selection	$W = 185.00, p = 0.009$	$t(25.8) = 2.76, p = 0.011$
LLM complex setup vs. LLM teachers' selection	$W = 192.00, p = 0.015$	$t(26.9) = 2.61, p = 0.014$

Table S.4

Kruskal-Wallis Test Results for Mathematics Items' Item Parameters - CR Items

Parameter	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi^2(3) = 2.12, p = 0.548$	$F(3, 33) = 1.28, p = 0.298$
Difficulty	$\chi^2(3) = 0.79, p = 0.851$	$F(3, 33) = 0.68, p = 0.573$
Mean log resp. time	$\chi^2(3) = 1.38, p = 0.709$	$F(3, 33) = 0.26, p = 0.855$

Note. This analysis compared CR TIMSS (N = 14) vs. CR LLM-generated items (N = 23).

Table S.5*Kruskal-Wallis Test Results for Mathematics Items' Parameters - MC Items*

Parameter	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi^2(3) = 2.85, p = 0.416$	$F(3, 54) = 1.05, p = 0.377$
Difficulty	$\chi^2(3) = 7.62, p = 0.055$	$F(3, 54) = 2.57, p = 0.064$
Mean log resp. time	$\chi^2(3) = 8.53, p = 0.036$	$F(3, 54) = 3.07, p = 0.035$

Note. This analysis compared MC TIMSS (N = 33) vs. MC LLM-generated items (N = 25).

Table S.6*Wilcoxon Post-Hoc Test Results for Mathematics Items' Parameters - MC Items*

Comparison	Wilcoxon rank sum exact test	t-test
Difficulty		
TIMSS vs. LLM (all items)	$W = 588.00, p = 0.005$	$t(26.4) = 2.12, p = 0.043$
TIMSS vs. LLM simple setup	$W = 191.00, p = 0.053$	$t(7.1) = 1.35, p = 0.220$
TIMSS vs. LLM complex setup	$W = 184.00, p = 0.091$	$t(7.9) = 1.26, p = 0.243$
TIMSS vs. LLM teachers' selection	$W = 213.00, p = 0.049$	$t(8.3) = 1.27, p = 0.238$
LLM simple setup vs. LLM complex setup	$W = 31.00, p = 0.959$	$t(8.6) = -0.86, p = 0.412$
LLM simple setup vs. LLM teachers' selection	$W = 37.00, p = 0.963$	$t(11.5) = -0.51, p = 0.622$
LLM complex setup vs. LLM teachers' selection	$W = 37.00, p = 0.963$	$t(12.5) = 0.47, p = 0.646$

Comparison	Wilcoxon rank sum exact	
	test	t-test
Mean log resp. time		
TIMSS vs. LLM (all items)	$W = 559.00, p = 0.021$	$t(49.4) = 2.21, p = 0.032$
TIMSS vs. LLM simple setup	$W = 140.00, p = 0.809$	$t(9.7) = 0.05, p = 0.958$
TIMSS vs. LLM complex setup	$W = 198.00, p = 0.029$	$t(10.1) = 2.31, p = 0.043$
TIMSS vs. LLM teachers' selection	$W = 221.00, p = 0.025$	$t(13.6) = 2.35, p = 0.035$
LLM simple setup vs. LLM complex setup	$W = 47.00, p = 0.130$	$t(13.9) = 1.69, p = 0.112$
LLM simple setup vs. LLM teachers' selection	$W = 53.00, p = 0.114$	$t(13.4) = 1.58, p = 0.138$
LLM complex setup vs. LLM teachers' selection	$W = 36.00, p = 1.000$	$t(13.9) = -0.29, p = 0.778$

Table S.7

Kruskal-Wallis Test Results for Science Items' Parameters - CR Items

Parameter	Kruskal-Wallis rank sum	
	test	ANOVA
Discrimination	$\chi^2(3) = 0.72, p = 0.869$	$F(3, 34) = 0.31, p = 0.818$
Difficulty	$\chi^2(3) = 15.44, p = 0.001$	$F(3, 34) = 7.08, p = 0.001$
Mean log resp. time	$\chi^2(3) = 8.61, p = 0.035$	$F(3, 34) = 3.38, p = 0.029$

Note. This analysis compared TIMSS (N=23) vs. LLM-generated items (N=15).

Table S.8*Kruskal-Wallis Test Results for Science Items' Parameters - MC Items*

Parameter	Kruskal-Wallis rank sum	
	test	ANOVA
Discrimination	$\chi^2(3) = 1.76, p = 0.624$	$F(3, 53) = 0.58, p = 0.629$
Difficulty	$\chi^2(3) = 0.89, p = 0.827$	$F(3, 53) = 2.62, p = 0.060$
Mean log resp. time	$\chi^2(3) = 8.30, p = 0.040$	$F(3, 53) = 3.54, p = 0.021$

Note. This analysis compared TIMSS (N=25) vs. LLM-generated items (N=32).

Table S.9*Wilcoxon Post-Hoc Test Results for Science Items' Parameters - CR Items*

Comparison	Wilcoxon rank sum exact	
	test	t-test
Difficulty		
TIMSS vs. LLM (all items)	$W = 71.00, p = 0.002$	$t(21.5) = -3.09, p = 0.005$
TIMSS vs. LLM simple setup	$W = 2.00, p = 0.000$	$t(10.6) = -5.56, p = 0.000$
TIMSS vs. LLM complex setup	$W = 36.00, p = 0.029$	$t(7.2) = -2.08, p = 0.075$
TIMSS vs. LLM teachers' selection	$W = 33.00, p = 0.373$	$t(1.1) = 0.63, p = 0.635$
LLM simple setup vs. LLM complex setup	$W = 21.00, p = 1.000$	$t(8.2) = 0.75, p = 0.472$
LLM simple setup vs. LLM teachers' selection	$W = 12.00, p = 0.071$	$t(1.2) = 2.57, p = 0.203$
LLM complex setup vs. LLM teachers' selection	$W = 12.00, p = 0.222$	$t(2.1) = 1.76, p = 0.217$
Mean log resp. time		
TIMSS vs. LLM (all items)	$W = 79.00, p = 0.004$	$t(22.2) = -2.82, p = 0.010$
TIMSS vs. LLM simple setup	$W = 24.00, p = 0.014$	$t(6.5) = -2.32, p = 0.057$
TIMSS vs. LLM complex setup	$W = 39.00, p = 0.042$	$t(7.1) = -1.87, p = 0.103$

Comparison	Wilcoxon rank sum exact	
	test	t-test
TIMSS vs. LLM teachers' selection	$W = 16.00, p = 0.540$	$t(1.4) = -0.80, p = 0.539$
LLM simple setup vs. LLM complex setup	$W = 22.00, p = 0.945$	$t(10.9) = 0.05, p = 0.959$
LLM simple setup vs. LLM teachers' selection	$W = 9.00, p = 0.429$	$t(4) = 1.30, p = 0.263$
LLM complex setup vs. LLM teachers' selection	$W = 11.00, p = 0.333$	$t(5.4) = 1.09, p = 0.320$

Table S.10

Wilcoxon Post-Hoc Test Results for Science Items' Parameters - MC Items

Comparison	Wilcoxon rank sum exact	
	test	t-test
Mean log resp. time		
TIMSS vs. LLM (all items)	$W = 537.00, p = 0.027$	$t(50.2) = 2.55, p = 0.014$
TIMSS vs. LLM simple setup	$W = 124.00, p = 0.673$	$t(13.1) = 0.60, p = 0.556$
TIMSS vs. LLM complex setup	$W = 144.00, p = 0.231$	$t(16.7) = 1.52, p = 0.148$
TIMSS vs. LLM teachers' selection	$W = 269.00, p = 0.005$	$t(31.8) = 3.43, p = 0.002$
LLM simple setup vs. LLM complex setup	$W = 47.00, p = 0.605$	$t(15) = 0.59, p = 0.565$
LLM simple setup vs. LLM teachers' selection	$W = 91.00, p = 0.083$	$t(13.7) = 1.79, p = 0.095$
LLM complex setup vs. LLM teachers' selection	$W = 83.00, p = 0.224$	$t(16.8) = 1.38, p = 0.185$

7.3 Additional results for DIF analysis

Table S.11*DIF Analysis for Mathematics Items*

Source	Item	Statistic	P-value	Holm's corrected p-value	Δ Nagelkerke's R ²	Effect size (Jodoin & Gierl)	More difficult for	More discriminating for
LLM	M090048	1.43	0.489	1.000	0.107	N/A	N/A	N/A
TIMSS	M041135	12.53	0.002	0.181	0.043	N/A	N/A	N/A
LLM	M890010	3.80	0.149	1.000	0.034	N/A	N/A	N/A
LLM	M890004	2.89	0.235	1.000	0.034	N/A	N/A	N/A
LLM	M890005	11.32	0.003	0.324	0.033	N/A	N/A	N/A
TIMSS	M061248	11.59	0.003	0.286	0.028	N/A	N/A	N/A
LLM	M090000	7.30	0.026	1.000	0.028	N/A	N/A	N/A
LLM	M100020	3.05	0.217	1.000	0.026	N/A	N/A	N/A
TIMSS	M051061	10.02	0.007	0.614	0.026	N/A	N/A	N/A
LLM	M090042	5.37	0.068	1.000	0.025	N/A	N/A	N/A
TIMSS	M041333	8.70	0.013	1.000	0.024	N/A	N/A	N/A
LLM	M090016	8.18	0.017	1.000	0.024	N/A	N/A	N/A

TIMSS	M041271	8.59	0.014	1.000	0.022	N/A	N/A	N/A
TIMSS	MP71217	8.56	0.014	1.000	0.022	N/A	N/A	N/A
TIMSS	M061167	7.06	0.029	1.000	0.019	N/A	N/A	N/A
LLM	M090007	5.06	0.080	1.000	0.018	N/A	N/A	N/A
TIMSS	M061106	6.21	0.045	1.000	0.018	N/A	N/A	N/A
LLM	M100010	5.12	0.077	1.000	0.014	N/A	N/A	N/A
TIMSS	M041059	5.81	0.055	1.000	0.014	N/A	N/A	N/A
LLM	M890039	1.97	0.373	1.000	0.014	N/A	N/A	N/A
TIMSS	M051205	4.50	0.105	1.000	0.013	N/A	N/A	N/A
LLM	M890029	2.77	0.250	1.000	0.013	N/A	N/A	N/A
TIMSS	M061236	3.51	0.173	1.000	0.012	N/A	N/A	N/A
LLM	M100028	2.25	0.324	1.000	0.011	N/A	N/A	N/A
LLM	M100009	2.98	0.225	1.000	0.011	N/A	N/A	N/A
TIMSS	M041048	3.82	0.148	1.000	0.011	N/A	N/A	N/A
TIMSS	M051236	2.56	0.278	1.000	0.011	N/A	N/A	N/A

LLM	M890003	3.49	0.175	1.000	0.010	N/A	N/A	N/A
TIMSS	M051039	3.64	0.162	1.000	0.010	N/A	N/A	N/A
LLM	M890024	1.69	0.429	1.000	0.010	N/A	N/A	N/A
TIMSS	M061207	3.73	0.155	1.000	0.010	N/A	N/A	N/A
TIMSS	M041177	2.29	0.318	1.000	0.009	N/A	N/A	N/A
LLM	M090047	2.87	0.238	1.000	0.008	N/A	N/A	N/A
TIMSS	M051077	3.02	0.221	1.000	0.008	N/A	N/A	N/A
TIMSS	M051070	2.71	0.258	1.000	0.008	N/A	N/A	N/A
LLM	M090057	2.68	0.261	1.000	0.008	N/A	N/A	N/A
LLM	M890011	1.91	0.384	1.000	0.007	N/A	N/A	N/A
TIMSS	M061039	2.63	0.268	1.000	0.006	N/A	N/A	N/A
TIMSS	M041268	2.27	0.321	1.000	0.006	N/A	N/A	N/A
TIMSS	M041174	2.16	0.339	1.000	0.006	N/A	N/A	N/A
LLM	M890006	2.02	0.364	1.000	0.006	N/A	N/A	N/A
TIMSS	M061052	1.95	0.377	1.000	0.006	N/A	N/A	N/A

TIMSS	M061050	1.95	0.377	1.000	0.006	N/A	N/A	N/A
LLM	M890025	1.83	0.400	1.000	0.005	N/A	N/A	N/A
LLM	M890021	1.16	0.560	1.000	0.005	N/A	N/A	N/A
LLM	M890030	1.56	0.459	1.000	0.005	N/A	N/A	N/A
TIMSS	M061274	1.49	0.475	1.000	0.004	N/A	N/A	N/A
LLM	M090013	1.53	0.466	1.000	0.004	N/A	N/A	N/A
TIMSS	M061179	1.42	0.491	1.000	0.004	N/A	N/A	N/A
TIMSS	M041169	1.50	0.473	1.000	0.004	N/A	N/A	N/A
LLM	M890002	1.20	0.550	1.000	0.004	N/A	N/A	N/A
TIMSS	M041298	0.97	0.616	1.000	0.004	N/A	N/A	N/A
LLM	M090017	0.62	0.733	1.000	0.004	N/A	N/A	N/A
LLM	M090012	0.58	0.750	1.000	0.004	N/A	N/A	N/A
LLM	M100017	0.81	0.666	1.000	0.004	N/A	N/A	N/A
TIMSS	M041151	1.25	0.536	1.000	0.003	N/A	N/A	N/A
LLM	M100008	0.96	0.618	1.000	0.003	N/A	N/A	N/A

LLM	M090059	0.85	0.655	1.000	0.003	N/A	N/A	N/A
TIMSS	M041034	1.16	0.560	1.000	0.003	N/A	N/A	N/A
TIMSS	M061206	1.10	0.576	1.000	0.003	N/A	N/A	N/A
TIMSS	M041280	1.03	0.596	1.000	0.003	N/A	N/A	N/A
LLM	M090044	0.69	0.708	1.000	0.003	N/A	N/A	N/A
TIMSS	M061272	0.99	0.610	1.000	0.003	N/A	N/A	N/A
LLM	M890026	0.86	0.651	1.000	0.002	N/A	N/A	N/A
TIMSS	M041191	0.57	0.754	1.000	0.002	N/A	N/A	N/A
LLM	M100011	0.65	0.724	1.000	0.002	N/A	N/A	N/A
LLM	M890009	0.78	0.678	1.000	0.002	N/A	N/A	N/A
LLM	M090055	0.34	0.845	1.000	0.002	N/A	N/A	N/A
LLM	M100027	0.49	0.783	1.000	0.002	N/A	N/A	N/A
TIMSS	MP71021	0.63	0.730	1.000	0.002	N/A	N/A	N/A
LLM	M100037	0.38	0.827	1.000	0.002	N/A	N/A	N/A
TIMSS	M051129	0.52	0.769	1.000	0.002	N/A	N/A	N/A

TIMSS	MP71219	0.47	0.792	1.000	0.001	N/A	N/A	N/A
TIMSS	MP71041	0.45	0.797	1.000	0.001	N/A	N/A	N/A
LLM	M100013	0.55	0.758	1.000	0.001	N/A	N/A	N/A
TIMSS	MP71162	0.44	0.801	1.000	0.001	N/A	N/A	N/A
LLM	M090051	0.48	0.785	1.000	0.001	N/A	N/A	N/A
LLM	M100026	0.44	0.802	1.000	0.001	N/A	N/A	N/A
LLM	M100039	0.44	0.802	1.000	0.001	N/A	N/A	N/A
TIMSS	MP71090	0.40	0.817	1.000	0.001	N/A	N/A	N/A
TIMSS	M041264	0.42	0.812	1.000	0.001	N/A	N/A	N/A
TIMSS	M061185	0.36	0.835	1.000	0.001	N/A	N/A	N/A
LLM	M090006	0.13	0.937	1.000	0.001	N/A	N/A	N/A
TIMSS	M041096	0.30	0.860	1.000	0.001	N/A	N/A	N/A
LLM	M890022	0.26	0.877	1.000	0.001	N/A	N/A	N/A
LLM	M100024	0.26	0.880	1.000	0.001	N/A	N/A	N/A
TIMSS	M061031	0.24	0.888	1.000	0.001	N/A	N/A	N/A

LLM	M100004	0.10	0.951	1.000	0.000	N/A	N/A	N/A
TIMSS	M051093	0.08	0.961	1.000	0.000	N/A	N/A	N/A
TIMSS	M041125	0.06	0.969	1.000	0.000	N/A	N/A	N/A
TIMSS	M041276	0.06	0.972	1.000	0.000	N/A	N/A	N/A
LLM	M090052	0.04	0.978	1.000	0.000	N/A	N/A	N/A
LLM	M100007	0.04	0.979	1.000	0.000	N/A	N/A	N/A
TIMSS	M061029	0.04	0.981	1.000	0.000	N/A	N/A	N/A
LLM	M100002	0.02	0.990	1.000	0.000	N/A	N/A	N/A

Table S.12

DIF Analysis for Science Items

Source	Item	Statistic	P-value	Holm's corrected p-value	Δ Nagelkerke's R2	Effect size (Jodoin & Gierl)	More difficult for	More discriminating for
LLM	S890009	19.69	0.000	0.005	0.058	moderate	males	males
LLM	S090051	17.99	0.000	0.012	0.051	moderate	females	males
TIMSS	S051121	17.64	0.000	0.014	0.051	moderate	females	males

TIMSS	S061031	17.06	0.00 0	0.018	0.049	moderate	females	males
LLM	S100032	12.31	0.00 2	0.193	0.105	N/A	N/A	N/A
LLM	S890017	9.81	0.00 7	0.637	0.084	N/A	N/A	N/A
TIMSS	S041033	11.88	0.00 3	0.237	0.049	N/A	N/A	N/A
LLM	S890038	11.19	0.00 4	0.330	0.048	N/A	N/A	N/A
TIMSS	S041010	10.43	0.00 5	0.477	0.045	N/A	N/A	N/A
LLM	S100010	3.11	0.21 1	1.000	0.036	N/A	N/A	N/A
TIMSS	S041183	7.93	0.01 9	1.000	0.032	N/A	N/A	N/A
LLM	S090052	3.76	0.15 3	1.000	0.031	N/A	N/A	N/A
LLM	S090018	9.45	0.00 9	0.754	0.029	N/A	N/A	N/A
TIMSS	S061138	10.12	0.00 6	0.552	0.028	N/A	N/A	N/A
TIMSS	S061105	8.85	0.01 2	0.996	0.026	N/A	N/A	N/A
LLM	S090080	7.15	0.02 8	1.000	0.026	N/A	N/A	N/A

LLM	S100014	6.18	0.04 5	1.000	0.025	N/A	N/A	N/A
TIMSS	S041081	8.94	0.01 1	0.964	0.025	N/A	N/A	N/A
LLM	S090041	6.67	0.03 6	1.000	0.024	N/A	N/A	N/A
LLM	S090077	8.48	0.01 4	1.000	0.024	N/A	N/A	N/A
LLM	S100024	5.70	0.05 8	1.000	0.023	N/A	N/A	N/A
TIMSS	S041027	2.67	0.26 3	1.000	0.023	N/A	N/A	N/A
TIMSS	S051105	6.88	0.03 2	1.000	0.022	N/A	N/A	N/A
TIMSS	SP61011	7.47	0.02 4	1.000	0.021	N/A	N/A	N/A
TIMSS	SP71147	7.14	0.02 8	1.000	0.021	N/A	N/A	N/A
LLM	S890022	4.27	0.11 8	1.000	0.021	N/A	N/A	N/A
TIMSS	SP71268	6.83	0.03 3	1.000	0.019	N/A	N/A	N/A
TIMSS	S061071	4.90	0.08 6	1.000	0.017	N/A	N/A	N/A
LLM	S100004	1.14	0.56 6	1.000	0.016	N/A	N/A	N/A

TIMSS	S061034	4.87	0.08 8	1.000	0.015	N/A	N/A	N/A
TIMSS	SP71072	5.03	0.08 1	1.000	0.014	N/A	N/A	N/A
TIMSS	S051147	4.29	0.11 7	1.000	0.014	N/A	N/A	N/A
LLM	S100015	4.37	0.11 2	1.000	0.014	N/A	N/A	N/A
TIMSS	S061065	4.80	0.09 1	1.000	0.014	N/A	N/A	N/A
TIMSS	S051087	3.54	0.17 0	1.000	0.013	N/A	N/A	N/A
TIMSS	S041052	4.26	0.11 9	1.000	0.012	N/A	N/A	N/A
LLM	S890021	3.53	0.17 1	1.000	0.012	N/A	N/A	N/A
TIMSS	S041068	3.49	0.17 4	1.000	0.010	N/A	N/A	N/A
LLM	S890016	3.61	0.16 4	1.000	0.010	N/A	N/A	N/A
LLM	S890006	1.75	0.41 7	1.000	0.010	N/A	N/A	N/A
TIMSS	S051110	3.31	0.19 1	1.000	0.010	N/A	N/A	N/A
TIMSS	S041070	3.42	0.18 1	1.000	0.009	N/A	N/A	N/A

TIMSS	S061028	2.43	0.29 7	1.000	0.009	N/A	N/A	N/A
TIMSS	SP71017	2.75	0.25 3	1.000	0.008	N/A	N/A	N/A
TIMSS	S051188	2.69	0.26 1	1.000	0.008	N/A	N/A	N/A
TIMSS	S041032	1.94	0.38 0	1.000	0.008	N/A	N/A	N/A
TIMSS	S051037	2.54	0.28 1	1.000	0.007	N/A	N/A	N/A
LLM	S100008	2.00	0.36 9	1.000	0.007	N/A	N/A	N/A
TIMSS	S041006	2.49	0.28 8	1.000	0.007	N/A	N/A	N/A
LLM	S890023	2.37	0.30 6	1.000	0.007	N/A	N/A	N/A
TIMSS	S041043	2.17	0.33 7	1.000	0.006	N/A	N/A	N/A
TIMSS	SP71054	1.93	0.38 2	1.000	0.006	N/A	N/A	N/A
LLM	S090048	1.96	0.37 6	1.000	0.006	N/A	N/A	N/A
TIMSS	SP71115	1.85	0.39 7	1.000	0.005	N/A	N/A	N/A
LLM	S090058	1.33	0.51 5	1.000	0.005	N/A	N/A	N/A

LLM	S090040	1.04	0.59 6	1.000	0.005	N/A	N/A	N/A
LLM	S090020	1.72	0.42 2	1.000	0.005	N/A	N/A	N/A
TIMSS	SP71402	1.37	0.50 4	1.000	0.005	N/A	N/A	N/A
LLM	S090064	1.47	0.48 0	1.000	0.005	N/A	N/A	N/A
TIMSS	S041102	1.29	0.52 5	1.000	0.005	N/A	N/A	N/A
LLM	S100029	1.34	0.51 2	1.000	0.004	N/A	N/A	N/A
TIMSS	S061060	1.42	0.49 2	1.000	0.004	N/A	N/A	N/A
LLM	S890002	0.68	0.71 2	1.000	0.004	N/A	N/A	N/A
LLM	S890005	0.42	0.81 0	1.000	0.004	N/A	N/A	N/A
LLM	S100016	1.25	0.53 4	1.000	0.004	N/A	N/A	N/A
TIMSS	S051111	1.22	0.54 4	1.000	0.004	N/A	N/A	N/A
TIMSS	SP71140	1.11	0.57 5	1.000	0.004	N/A	N/A	N/A
TIMSS	S061098	1.17	0.55 7	1.000	0.003	N/A	N/A	N/A

LLM	S890011	1.25	0.53 6	1.000	0.003	N/A	N/A	N/A
TIMSS	S041107	0.77	0.68 0	1.000	0.003	N/A	N/A	N/A
TIMSS	SP61044	0.97	0.61 4	1.000	0.003	N/A	N/A	N/A
TIMSS	SP71128	0.68	0.71 0	1.000	0.002	N/A	N/A	N/A
TIMSS	S061130	0.64	0.72 6	1.000	0.002	N/A	N/A	N/A
LLM	S100009	0.63	0.73 0	1.000	0.002	N/A	N/A	N/A
TIMSS	S061010	0.80	0.66 9	1.000	0.002	N/A	N/A	N/A
LLM	S100012	0.52	0.77 0	1.000	0.002	N/A	N/A	N/A
LLM	S100002	0.53	0.76 6	1.000	0.002	N/A	N/A	N/A
TIMSS	S051164	0.24	0.88 7	1.000	0.002	N/A	N/A	N/A
LLM	S890007	0.30	0.86 3	1.000	0.001	N/A	N/A	N/A
LLM	S890010	0.48	0.78 7	1.000	0.001	N/A	N/A	N/A
LLM	S100039	0.30	0.85 9	1.000	0.001	N/A	N/A	N/A

LLM	S890026	0.39	0.82 4	1.000	0.001	N/A	N/A	N/A
TIMSS	S051114	0.40	0.82 0	1.000	0.001	N/A	N/A	N/A
LLM	S890004	0.31	0.85 8	1.000	0.001	N/A	N/A	N/A
LLM	S090084	0.34	0.84 4	1.000	0.001	N/A	N/A	N/A
TIMSS	S041209	0.21	0.89 8	1.000	0.001	N/A	N/A	N/A
LLM	S100023	0.11	0.94 9	1.000	0.001	N/A	N/A	N/A
LLM	S090075	0.15	0.92 6	1.000	0.001	N/A	N/A	N/A
LLM	S090015	0.12	0.94 1	1.000	0.000	N/A	N/A	N/A
TIMSS	SP71077	0.08	0.96 2	1.000	0.000	N/A	N/A	N/A
LLM	S090053	0.07	0.96 6	1.000	0.000	N/A	N/A	N/A
LLM	S090001	0.04	0.97 8	1.000	0.000	N/A	N/A	N/A
TIMSS	S051201	0.04	0.98 1	1.000	0.000	N/A	N/A	N/A
LLM	S890024	0.04	0.98 1	1.000	0.000	N/A	N/A	N/A

LLM	S100020	0.02	0.99 2	1.000	0.000	N/A	N/A	N/A
-----	---------	------	-----------	-------	-------	-----	-----	-----

7.4 List of Annexes

The report is accompanied by five online Annexes:

- Annex 1 - presents details of psychometric analyses and provides a comprehensive psychometric profile of each of the items that were used in the study,
- Annex 2 - contains all of the items, along with their translations to English, used in the validation study,
- Annex 3 - contains the items generated as the part of the Subject Matter Experts (SMEs) study in their raw form, along with their translations to English and associated images and tables,
- Annex 4 - contains examples of generated items that were deemed unacceptable from among those generated for the validation study and were eliminated from the study.
- Annex 5 - contains the item-scoring key pairs used as examples in the scoring key generation process.