Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: www.tandfonline.com/journals/hame20

Establishing Cognitive Item Models for Fair and Theory-Grounded Automatic Item Generation: A Large-Scale Assessment Study with Image-Based Math Items

Philipp Sonnleitner, Steve Bernard, Michael A. Michels, Pamela Inostroza-Fernandez, Ulrich Keller, Mark J. Gierl, Pedro Cardoso-Leite & Caroline Hornung

To cite this article: Philipp Sonnleitner, Steve Bernard, Michael A. Michels, Pamela Inostroza-Fernandez, Ulrich Keller, Mark J. Gierl, Pedro Cardoso-Leite & Caroline Hornung (14 Nov 2025): Establishing Cognitive Item Models for Fair and Theory-Grounded Automatic Item Generation: A Large-Scale Assessment Study with Image-Based Math Items, Applied Measurement in Education, DOI: 10.1080/08957347.2025.2563889

To link to this article: https://doi.org/10.1080/08957347.2025.2563889

Published online: 14 Nov 2025.

APPLIED MEASUREMENT IN EDUCATION
https://doi.org/10.1080/08957347.2025.2563889

Routledge
Taylor & Francis Group

Establishing Cognitive Item Models for Fair and Theory-Grounded Automatic Item Generation: A Large-Scale Assessment Study with Image-Based Math Items

Philipp Sonnleitner[a], Steve Bernard[a], Michael A. Michels[a], Pamela Inostroza-Fernandez[b], Ulrich Keller[a], Mark J. Gierl[c], Pedro Cardoso-Leite[d], and Caroline Hornung[a]

[a]Luxembourg Centre for Educational Testing, University of Luxembourg; [b]Innovation Department, Universidad de los Andes; [c]Faculty of Education, University of Alberta; [d]Department of Behavioural and Cognitive Sciences, University of Luxembourg

ABSTRACT

Mathematics is a core domain in large-scale assessments (LSA), yet item development remains resource-intensive, limiting scalability and innovation. Automatic Item Generation (AIG) offers a promising solution, but empirical validations remain rare. This study investigates the psychometric functioning and fairness of 48 cognitive item models designed to generate language-reduced, image-based math items for Grades 1, 3, and 5. Treating these models as proto-theories, we generated 612 item instances varying in cognitive demands and contextual features. Using data from Luxembourg's school monitoring ($N = 35{,}058$), we found that item difficulty was mainly driven by predefined cognitive factors, with stronger contextual influences in early grades. We introduce *Differential Radical Functioning* to evaluate whether AIG-based items permit comparable score interpretations across subgroups. Results reveal meaningful differences by cultural background, regardless of language proficiency. These findings highlight the importance of contextual embedding and demonstrate the potential of cognitive modeling in AIG for scalable, valid, and equitable assessments.

1. Introduction

Mathematical skills are fundamental in today's society and serve as strong predictors of key life outcomes and broader competencies (Davis-Kean et al., 2022; Sells, 1973). As a consequence, mathematics remains one of the core competency domains assessed regularly by large-scale educational assessments (LSAs) to evaluate and monitor programs and curricula. These include international studies such as TIMSS and PISA, as well as national assessments like the U.S. NAEP (De Lange, 2007; National Center for Education Statistics, 2022; OECD, 2024; von Davier et al., 2024). What follows is a huge and constant demand of valid test items, especially for national assessments often covering a broad range of K-12 mathematical learning trajectories. Those items are typically a) developed in small groups of subject matter experts (SMEs) and psychometricians, b) prepared for the respective mode of test administration (paper-pencil or computer-based), c) extensively reviewed, and finally d) field-tested to gather empirical evidence on reliability and validity – a process that affords a huge amount of human and technical resources (Harris et al., 2024; Kosh et al., 2019).

CONTACT Philipp Sonnleitner philipp.sonnleitner@uni.lu Luxembourg Centre for Educational Testing, University of Luxembourg, 11, Porte des Sciences, Esch-sur-Alzette L-4366, Luxembourg

A potential solution to this is template-based automatic item generation (AIG), an approach that uses prespecified item models which are translated by computer algorithms into items (Gierl & Haladyna, 2013; S. Irvine & Kyllonen, 2002). Because of its scalability, AIG reduces the cost of individual items, increases test security by limiting the exposure of individual items, and, if the underlying item models are based on cognitive theory, can offer additional validity evidence and enable the prediction of psychometric characteristics without the need for expensive field tests (Embretson & Kingston, 2018; Gierl & Lai, 2016). However, operational implementation of AIG remains rare (for notable exceptions see Attali, 2018; Gierl et al., 2015; Sinharay & Johnson, 2008), even more so with cognitive theories as a foundation for item model development. This scarcity may stem from the substantial initial effort required to implement AIG or from testing companies withholding disclosure, likely due to concerns over intellectual property (Circi et al., 2023; Harris et al., 2024; Kosh et al., 2019). As a consequence though, it is hard to judge whether AIG would indeed offer advantages over traditional item development for LSAs, especially concerning an increased validity argument. Further, almost all use cases generate text-based stimuli and responses, making reading comprehension crucial by potentially introducing construct-irrelevant variance, especially in today's increasingly multilingual student populations (Greisen et al., 2021). Whether generating image-based items instead of text-based ones might mitigate this issue isn't known and differences based on cultural background knowledge might persist.

The current paper addresses these limitations by exploring the implementation of AIG in a national mathematics LSA which relies heavily on language-reduced, image-based items. We developed 48 cognitive item models to generate 612 item instances for two mathematical domains: *Numbers & Operations (NO)* and *Space & Shape (SS)*. Validity and fairness of these items with regard to students' background characteristics and math-related attitudes were studied in Grades 1, 3, and 5. Results show that item difficulty parameters could be explained to a large degree by model-specific cognitive and contextual factors, and that subgroup differences based on different contexts vanished with increasing students' age.

1.1. Automatic Item Generation in Mathematics and Its Validity

Due to its clear logic and relations between its components, mathematics has been a promising candidate for AIG from the start. For the competency range that is relevant for educational LSAs (K-12), many mathematics item generators have been developed since then. Examples include the Mathematics Test Creation Assistant (MTCA; Singley & Bennett, 2002), IGOR (Gierl et al., 2008), MathGen (Wilson et al., 2014), Quick Math (Attali & Arieli-Attali, 2015), and Mathematics Item Generator (MIG; Kellogg et al., 2015).

Early AIG efforts often relied on theoretical cognitive models to guide item generation and explain or predict item characteristics – a strategy known as "strong theory" approach (e.g. Drasgow et al., 2006). Over time, however, the focus shifted to a more pragmatic "weak theory" approach, which prioritizes the volume of generated items and technological efficiency. State-of-the-art edited volumes on AIG mirror this trend: S. Irvine and Kyllonen (2002), as well as Gierl and Haladyna (2013) feature a handful of chapters on explanation or prediction of psychometric item characteristics, whereas Gierl et al. (2021) mostly focus on increasingly sophisticated techniques for item model development and item banking. This shift has good reasons. Incorporating cognitive theories is costly, requiring significant expertise and time to translate them into item blueprints. In most applied settings, it is often sufficient to generate a large number of item instances based on proven parent items (which can be considered item clones) or to validate items using conventional methods, such as expert reviews or field tests. Most importantly though, for many competencies, including mathematics, robust and comprehensive cognitive theories supporting a strong theory-driven approach are either lacking, underdeveloped, or overly fine-grained (Leighton & Gierl, 2011). This mismatch in granularity is even more pronounced for competencies typically assessed in LSAs, which often extend beyond basic arithmetic skills to encompass applied problem-solving or are, in general, more oriented toward

didactical frameworks (e.g. De Lange, 2007). Ideally, a cognitive theory would provide a clear basis for understanding the thought processes involved in solving specific math problems. Such a framework would enable predictions about which problem aspects might influence the solving process and, therefore, merit assessment. These aspects, when translated into item features, are referred to as "radicals" (Gierl et al., 2008; S. H. Irvine, 2002). Conversely, aspects that do not affect item difficulty, known as "incidentals," can be varied in problems without altering their difficulty. Only when such cognitive theories are available and can be applied to the development of item models – enabling a strong theory-driven approach – do immediate benefits for test validity arise. These include defensible construct validity at the item level, as well as the ability to explain and potentially predict item characteristics for the generated items (Embretson & Kingston, 2018; Gierl & Lai, 2016).

As indicated above, examples of a strong theory approach are generally scarce and for the domain of mathematics are mostly restricted to word problems of complex quantitative reasoning, number learning, and algebra at middle school to university entry level (e.g. Arendasy & Sommer, 2007; Attali, 2018; Daniel & Embretson, 2010; Embretson & Kingston, 2018). In the context of psychometric modeling of generated items, these efforts proved highly beneficial though. Strong theory-driven approaches facilitated explanatory modeling of item difficulty and discrimination parameters. In contrast, weak theory approaches often struggled to produce items with consistent characteristics – so-called isomorphs – even within item families that were intended to share the same theoretical difficulty. This variation necessitated the inclusion of random effects at the item level during modeling (Attali, 2018; Cho et al., 2014; Sinharay & Johnson, 2008).

In conclusion, anchoring AIG in theoretical cognitive models offers clear benefits but comes with significant challenges. Chief among these are the substantial effort required and the limited availability of suitable cognitive theories. To bridge this gap, it may be useful to adopt the concept of "proto-theories" (e.g. Borsboom et al., 2021; Nettle, 2021). Proto-theories compile research findings on isolated phenomena – such as specific aspects of mathematical cognition – without committing to overly strong assumptions about the scope or magnitude of these effects. These proto-theories could provide mechanistic insights into certain problem-solving processes, serving as a foundation for exploratory, hypothesis-driven experimentation in item development. By taking a more flexible and iterative approach, researchers could identify key "levers" that influence item parameters, thereby gradually mapping out a competency domain. This approach could bridge the gap between cognitive theory and item development while advancing cognitive science. Targeting underexplored mathematical domains, especially in elementary education, where research results are plenty but AIG applications are sparse, could be highly impactful. Despite initial substantial investment, this approach offers lasting benefits for continued LSA programs.

1.2. Automatic Item Generation in Mathematics and Its Fairness

One of the major challenges in today's diverse educational systems is the growing number of students with migration backgrounds. These students often bring multilingual and multicultural experiences but frequently lack proficiency in the language of instruction (e.g. OECD, 2012). This language barrier directly translates into weaker performance in mathematics, particularly in text-based math problems, as several studies have demonstrated (Greisen et al., 2021; Saalbach et al., 2016; Singer & Strasser, 2017). However, the challenge extends beyond language; immigrant students also bring different sociocultural knowledge compared to their native peers (Martin et al., 2012; Sam et al., 2022).

In mathematics LSAs, item formats vary in the degree to which they rely on linguistic input and can be broadly classified as numeric (purely symbolic), text-based (verbal problem statements), and language-reduced/image-based (using primarily visual representations to convey context) (cf. Nohara, 2001). Language-reduced items aim to minimize the influence of reading comprehension by replacing textual contexts with images or diagrams, a strategy adopted in multilingual contexts such as Luxembourg's ÉpStan (Sonnleitner et al., 2018). While visuals can mitigate construct-irrelevant variance due to language (Greisen et al., 2021; Singer & Strasser, 2017), they may still introduce

sociocultural biases through the choice of depicted contexts (Martin et al., 2012; Sam et al., 2022), particularly when problems are embedded in real-world scenarios. Contextualized math problems – those that mimic real-life situations – are prevalent in LSAs like TIMSS-R (44% of problems), NAEP (48%), and PISA (97%) (Nohara, 2001). These problems require students to interpret everyday knowledge and translate it into a mathematical framework to arrive at a solution (Blum & Leiss, 2007).

The influence of such "cover stories" has been documented, with students doing better on problems featuring familiar or interesting contexts (Leiss et al., 2024). For example, De Lange (2007) reported that Dutch students performed better in a culturally adapted version of TIMSS, speculating that culturally appropriate embedding of anchor items improved their outcome. Similarly, a study on automatically generated algebra word problems found a weak but measurable effect of the typicality of the cover story on item difficulty (Arendasy & Sommer, 2007). Notably, in the published AIG literature, examples predominantly focus on the generation and variation of text-based items (e.g. Arendasy & Sommer, 2007; Gierl et al., 2008; Kellogg et al., 2015; Wilson et al., 2014). By contrast, language-reduced, image-based AIG has received far less attention in the open literature, and many commercial systems are not publicly documented. For AIG to be effective in contextual, language-reduced LSA settings, it must adopt similar approaches, replacing variations in word-based contexts with variations in illustrations while ensuring that these illustrations function as *incidentals* rather than *radic*als.

Research on the effects of visual aids in math problems presents mixed findings. While some studies suggest that visual aids in depictive representations of math problems have little to no impact on performance (Dewolf et al., 2017), others indicate that math problems presented visually are easier to solve than their purely textual counterparts (Hoogland et al., 2018). Beyond performance, illustrations can also influence test motivation. A study conducted within Luxembourg's school monitoring program *Épreuves Standardisées* (ÉpStan) revealed that illustrated contexts significantly impacted students' motivation (Girardelli, 2017). This finding is particularly relevant, as personal interest is a strong predictor of math problem-solving success (Walkington, 2013). However, given sociocultural differences in immigrant students' knowledge (Martin et al., 2012; Sam et al., 2022), culturally embedded illustrations may not have the same effect on all students and may inadvertently introduce cultural bias.

To account for such biases and to ensure the comparability of score interpretations across subgroups, educational LSAs routinely collect evidence for what the Standards for Educational and Psychological Testing refer to as interpretive fairness (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing, 2014). Traditional fairness approaches examine this comparability at the construct level through measurement invariance testing (e.g. Meredith, 1993) or at the item level through differential item functioning (DIF; e.g. Holland & Wainer, 2012; Wu et al., 2016; Zumbo, 1999). In the context of AIG, however, relying solely on post-administration DIF analyses undermines the potential benefits of the technology. A more proactive approach involves adapting DIF logic to the item-component level, assessing fairness before items are assembled.

Our study carries this principle down to the level of item components. Specifically, we investigate whether particular cognitive or contextual features of an item model systematically advantage or disadvantage certain subgroups, an approach we refer to as *differential radical functioning* (DRF). Unlike traditional DIF analyses, DRF examines potential sources of subgroup performance differences during item model design and validation. This proactive, component-level perspective embeds fairness considerations directly into the AIG process, providing early evidence to support interpretive fairness before items are assembled into full scales, and is consistent with the evolution toward third-generation DIF praxis (Zumbo, 2007), which emphasizes understanding the sources of subgroup differences and incorporating contextual variables into fairness evaluations. For example, DRFs in cognitively based radicals might indicate subgroup weaknesses that could inform targeted interventions, such as training students in decade-crossing addition. In contrast, DRFs in contextual factors

would reveal fundamental differences in students' understanding of the math problem, likely due to sociocultural differences.

1.3. Aim and Scope of the Study

To expand knowledge on and experiences with cognitively based AIG and its fairness within an operational LSA setting, the current study adopted a multifaceted approach. Situated within the Luxembourgish school monitoring program *Épreuves standardisées* (ÉpStan; Sonnleitner et al., 2018), we incorporated the ideas of proto-theories and DRF into item model development and its analyses. ÉpStan covers the whole elementary school mathematics curriculum starting with Grade 1 and including Grades 3, and 5. Similar to TIMSS and NAEP, about half of the items (40–50%) feature math problems situated in real-world settings. Due to the highly heterogeneous and multilingual student population of Luxembourg, items are language-reduced and mostly based on images which allows to explore related validity and fairness aspects. We purposefully adopted an exploratory stance in this study to maximize insights for future applications of AIG in mathematical LSAs. Within this exploratory framework, we addressed two research questions:

RQ1: Can the difficulty of AIG-generated, language-reduced math items be predicted and explained based on predefined cognitive item models, treated here as proto-theories?

RQ2: Do contextual embeddings in these items support interpretive fairness across student subgroups, or do certain contextual features systematically advantage or disadvantage specific groups?

2. Methods

2.1. Item Model Development and Implementation in R Code

The starting point for item model development was ÉpStan's mathematics item data base. Building on Gierl et al. (2015), we first identified so-called *parent items* that are ideal representations of the measured competency and that had shown good psychometric performance in prior administrations (i.e. acceptable infit and item total correlations, no DIF). Subject Matter Experts (SMEs; teachers trained in item and test development and cognitive psychologists) translated these parent items into cognitive item models. In this process, item components were classified either as radicals (influencing item difficulty) or as incidentals (features expected not to affect difficulty). Radicals were defined based on relevant cognitive psychological research on mathematical development. In domains or subcompetencies where the research base was sparse (especially within the *SS* domain), assumptions were primarily informed by teachers' classroom experience and curricular knowledge on learning progressions, which have been shown to be empirically promising (e.g. Attali & Arieli-Attali, 2019). This pragmatic approach allowed us to formulate cognitive proto-theories for each item model. Contextual embeddings were selected by SMEs, drawing on materials familiar from classroom use or prior ÉpStan tests. These embeddings were hypothesized to serve as incidentals, but their role was later examined empirically. To support fairness, SMEs also screened the chosen contexts for potentially biased or culturally sensitive material. Figure 1 illustrates an example of such an item model.

All models were implemented in R software. The R code allowed experimenters to (a) specify values of the radicals (e.g. number ranges, object sequences, order requirements), (b) select among alternative contexts, and (c) generate an arbitrary number of item instances by systematically varying these parameters. Items were exported in PDF format, making them directly compatible with ÉpStan's paper-based delivery.

In total, 48 item models were developed, covering Grades 1, 3, and 5 and being representative of the Luxembourgish elementary school mathematics curriculum; 24 models capturing

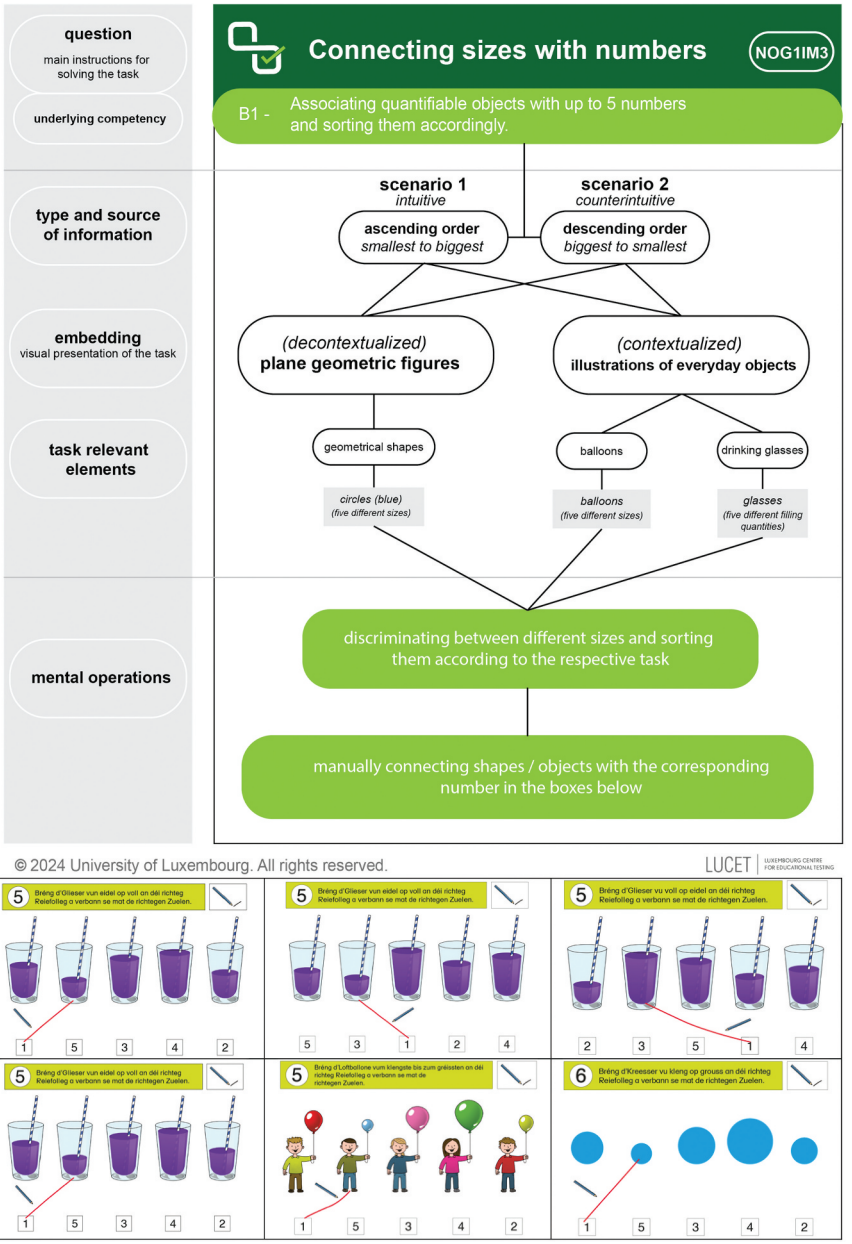


Figure 1. Item model for NOG1IM3 with six of the twelve generated and administered instances. The first row varies cognitive factors (radicals: inversion and sequence) while keeping the context constant, whereas the second row varies the contextual embedding (hypothesized incidentals) while maintaining identical cognitive characteristics.

the mathematical domain *NO* (8 item models per Grade), and 24 measuring the domain *SS* (again, 8 item models per Grade). Those models spanned six competency levels and covered 26 out of 42 subcompetencies judged relevant and testable in the context of a paper-and-pencil LSA. Annex 1 provides a schematic overview of this multi-step process, illustrating the role of SMEs at each stage, from parent item selection to implementation into the research design. Annex 2 and 3 summarize the developed item models, targeted subcompetencies, and generated item instances.

2.2. Research Design per Model

To investigate the item models' validity and fairness, each model was used to generate several item instances which incorporated different values of item radicals (e.g. number range or carry-over in an addition problem) that were expected to influence item difficulty. Item instances with identical problem characteristics (isomorphs) were then generated using different semantic contexts (e.g. contextualized vs. decontextualized). This design allowed for investigating a) validity of the theoretically supposed radicals, as well as b) fairness of the implemented different semantic contexts by psychometrically modeling and contrasting these aspects. Figure 1 illustrates the experimental design for item model NOG1IM3, showing six out of twelve generated and administered item instances. Two cognitive factors were hypothesized as radicals: the order of sorting (ascending/intuitive vs. descending/counterintuitive, termed inversion) and the sequence of objects, which was randomized. The radical inversion (i.e. sorting objects in descending rather than ascending order) was expected to increase item difficulty, as it requires children to inhibit the intuitive association between larger objects and larger numbers – a process known to challenge inhibitory control (e.g. Van Dooren & Inglis, 2015). To examine whether contextual embeddings functioned as incidentals, some instances kept the cognitive factors constant, while only the embedding was varied. Thus, inversion and sequence were hypothesized radicals, while embeddings were hypothesized incidentals, with their actual role examined empirically in the analyses.

Based on the 48 item models, a total of 360 item instances were generated within the domain of *NO*, and 252 for *SS*. See the Annex 2 and 3 for an overview of generated items per measured competency and model.

2.3. Data Collection

Data collection took place within the annual ÉpStan in November 2021 (with a focus on *NO*) and 2022 (focusing on *SS*). Full student cohorts were tested in Grades 1, 3, and 5. See Table 1 for sample descriptives. In addition to the mathematics main test, randomly drawn samples of students worked on one out of six different pretest booklets containing the generated items from the test design. Teachers administered the tests during class time.

2.4. Variables

2.4.1. Mathematical competency

Students' competencies in the domains *NO*, as well as *SS* were measured with the generated items based on the developed item models. The majority of items used a 1 out of 4 multiple-choice response format. In addition, some items requested the students to write their response in an open format or link elements of two lists of five objects. Responses were coded dichotomously by the administrating teachers using a web app provided by the school monitoring program.

2.4.2. Migration background

This variable was coded as binary: students were classified as native if both they and at least one parent were born in Luxembourg; all others were classified as having a migration background.

2.4.3. Language background

Based on previous studies conducted within Luxembourg's national school monitoring (Greisen et al., 2021; Sonnleitner et al., 2018), speaking Luxembourgish or German at home was considered advantageous for solving mathematics items in this context. Students reporting either Luxembourgish or German as a home language were classified as germanophone; all others were classified as non-germanophone.

Table 1. Sample descriptives for study 1 and study 2.

	n (Main test)	n (Pretests)						Age (years, Md)	Gender (% girls)	Migration background (% native)	Language background (% germanophone)	Interest in math; Mean (SD)	Math anxiety; Mean (SD)	Math self-concept; Mean (SD)
		1	2	3	4	5	6							
Numbers & Operations (ÉpStan 2021)														
Grade 1	5963	512	432	488	474	365	433	6	48.8%	37%	30.9%	3.32 (1.25)	1.66 (1.24)	3.41 (0.98)
Grade 3	5527	672	657	746	730	748	573	8	48.8%	40.3%	33.6%	3.25 (0.87)	1.35 (0.82)	3.23 (0.76)
Grade 5	5291	610	538	628	610	597	566	11	48.5%	38.8%	33.5%	3.19 (0.82)	1.25 (0.63)	3.25 (0.68)
Space & Shape (ÉpStan 2022)														
Grade 1	6302	608	622	572	652	619	621	6	49.1%	37.2%	31.4%	3.39 (1.20)	1.61 (1.21)	3.40 (0.97)
Grade 3	6212	718	798	779	813	818	699	8	48.3%	37.3%	32.4%	3.30 (0.85)	1.35 (0.8)	3.23 (0.75)
Grade 5	5763	625	575	682	578	636	620	11	48.7%	39%	33.4%	3.12 (0.85)	1.30 (0.69)	3.21 (0.69)

2.4.4. Socioeconomic status

Parents' current occupations were coded using the ISEI (International Socio-Economic Index of Occupational Status; Ganzeboom et al., 1992). The highest ISEI of either parent was used as indicator of the student's socioeconomic status (SES). Analyses compared the highest quartile (Q4) to the rest.

Interest in Mathematics, Math Self-Concept, and Math Anxiety. In Grade 1, each construct was measured by a positively phrased statement that students had to agree or disagree with. In Grades 3, and 5 *Interest in Mathematics* and *Math Self-Concept* were operationalized by two, *Math Anxiety* by one 4-point Likert-scale item(s). Binary responses of Grade 1 were coded as either 1 (disagree) or 4 (agree) to facilitate comparison between the Grades. Similar to SES, related analyses contrasted the highest quartile (Q4) to the rest.

2.5. Statistical Analyses

2.5.1. RQ1 – Validity of Item Models

After initial descriptive analyses, items from the main operational test, which were designed to assess the same competencies as those targeted by the item models, were screened using standard LSA quality criteria (cf. Wright & Masters, 1982; Wu et al., 2016): Items with corrected item-total correlations (rit) below .25, indicating weak discrimination, or weighted mean square infit statistics above 1.2, indicating misfit to the measurement model, were excluded from serving as anchors. This screening ensured that only reliable, well-fitting items contributed to linking the model-generated items across the six pretest booklets. Anchor items were estimated separately, and their parameters served as a common empirical reference for linking. All pretest items were retained regardless of these thresholds to allow evaluation of the full range of item model performance. Each item model was then analyzed separately drawing on a generalized linear mixed model in which each particular item is fixed, with students treated as random effects (GLMM De Boeck et al., 2011). To study validity of item models, we first estimated a fully saturated model as a benchmark ($model_0$), allowing one parameter for each individual item. This model was then contrasted to different reduced models with various cognitive and contextual factors acting as fixed effects rather than individual items following a stepwise, confirmatory procedure.

First, items were grouped based on cognitive characteristics of the respective model ($model_{cog}$) that were considered while creating the items (i.e. items sharing the same cognitive characteristics were modeled using one parameter). Second, this model was expanded by considering contextual factors ($model_{con}$; i.e. items sharing the same cognitive and contextual factors were modeled using one parameter). If the Likelihood Ratio Test (LRT) between the saturated and the reduced model was non-significant ($p > .05$), the more parsimonious, reduced model was preferred as explanatory model. If item difficulties could not be fully explained using this approach, additional exploratory reduced models ($model_{alt.x}$) were formulated and subsequently tested, partly revealing new cognitive or contextual factors not considered in the initial item model design. Items sharing identical characteristics with other item instances but not fitting the model were designated as outliers and assigned individual item parameters.

2.5.2. RQ2 – Fairness of Item Models

The final, most parsimonious model ($model_{fin}$), which was non-significant, served as the baseline for evaluating the fairness of the models with respect to subgroup differences. To detect potential fairness violations, we extended $model_{fin}$ by adding interaction terms between each item-level characteristic (cognitive and contextual item components) and the studied background variables (migration background, language background, socioeconomic status, interest in mathematics, math self-concept, and math anxiety). Significant interaction effects indicate that a particular item feature systematically advantaged or disadvantaged a specific subgroup, a pattern we refer to as Differential Radical Functioning (DRF, see above).

It is important to note that this approach focuses on item-component-level differences rather than general subgroup differences: main effects of background variables were not of interest here, as they may reflect broader construct-relevant performance gaps. Instead, DRF analyses specifically targeted whether certain cognitive or contextual features interacted with subgroup membership to produce construct-irrelevant variance in item difficulty.

All analyses were conducted using the R statistical framework (R Core team, 2024), using the packages TAM (Robitzsch et al., 2024) for item response modeling and lme4 (Bates et al., 2015) for mixed-effects modeling. Statistical significance was set at $p < .05$.

3. Results

3.1. RQ1 – Validity of Item Models

3.1.1. Explainability of Item Difficulty Variance by Cognitive Factors

3.1.1.1. Numbers & Operations. Of the 156 item instances administered in Grade 1, 80.76% could be modeled based on cognitive and/or contextual factors without the need for item-specific parameters. This indicates that the difficulty of the vast majority of items could be explained by these predefined and exploratory factors. Figure 2 shows the percentage of items accounted for by these factors across models (please see also Annex 2 and 3 for the number of item instances, number of parameters of $model_{\text{fin,}}$ and number of outliers per item model). Four item models could be fully explained by factors implemented in the design (100%; i.e. no outliers), while the model with the highest proportion of outliers explained only 45.83% of its items. The remaining four models revealed additional factors during the exploration process. Across all models, an average of 81.25% of items were successfully grouped ($SD = 21.36$). Table 2 provides a complete list of cognitive factors, indicating whether they were implemented by design or discovered during exploration, and whether they significantly impacted item difficulties.

The results for Grade 3 were slightly less promising: 70.17% of the 114 item instances were explained by cognitive and contextual factors ($M = 77.07\%$, $SD = 31.73$ across the 8 item models). Notably, item instances of one model (NOG3IM14) could not be grouped at all, requiring 24 individual parameters. However, the remaining 7 item models were fully explained by their design characteristics, with no need for further exploration. One of these models (NOG3IM5, comprising 18 items) was fully explained without any outliers.

Grade 5 showed remarkable 91.11% of the administered 90 item instances explained, mostly by cognitive factors that were implemented in the design ($M = 88.54\%$, $SD = 23.11$ across the 8 item models). Item instances of 5 item models were fully explained (two of them only needing 1 parameter each); only 1 item model revealed a factor through exploration.

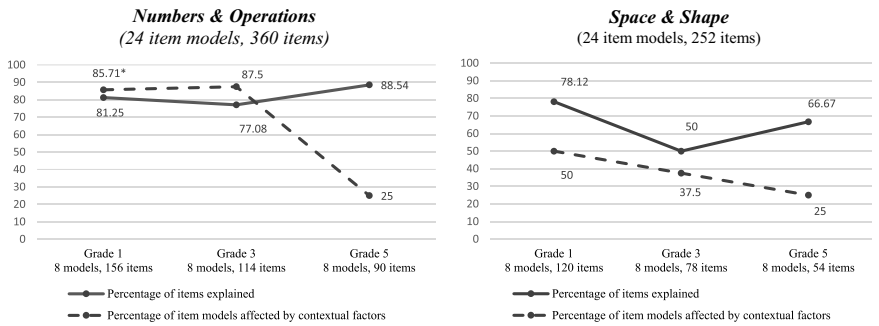


Figure 2. Mean percentage of items explained per model through cognitive and/or contextual factors (solid line) and percentage of item models significantly affected by contextual factors (dashed line). Results for grades 1, 3, and 5 are given for *Numbers & Operations* (left), and *Space & Shape* (right). *for *Numbers & Operation*, Grade 1, only seven item models were varied concerning contextual factors.

Table 2. Overview on studied cognitive factors in *Numbers & Operations* to explain item difficulty of instances.

Cognitive factor	Grade	IM	design	explored
Sequence of objects to compare	1	NOG1IM1	1*	
	1	NOG1IM2	1	
	1	NOG1IM3	1	
Orientation of objects to compare (horizontal vs vertical)	1	NOG1IM2		1*
Inversion	1	NOG1IM2	1*	
	1	NOG1IM3	1*	
	1	NOG1IM4	1*	
	3	NOG3IM4	1*	
	3	NOG3IM9	1	
Number range	1	NOG1IM4	1*	
	1	NOG1IM5	1*	
	1	NOG1IM6	1*	
	1	NOG1IM7	1*	
	5	NOG5IM11	1	
Tie effect ($r = s$)	1	NOG1IM5	1*	
	1	NOG1IM7	1*	
r > s	1	NOG1IM7	1*	
r < s	1	NOG1IM7	1*	
Even/odd difference in calculation	1	NOG1IM8		1*
Borrowing	3	NOG3IM8	1*	
	3	NOG3IM11	1*	
	5	NOG5IM11	1*	
Carry-over	3	NOG3IM10	1*	
	5	NOG5IM17	1	
Marking even/odd	3	NOG3IM14	1*	
Multiplication problem	3	NOG3IM11	1*	
Number of summands	3	NOG3IM13	1*	
	5	NOG5IM13	1*	
Different scaling when reading bar charts	5	NOG5IM15	1*	
Position of missing value when reading bar charts	5	NOG5IM15	1*	
Position of error to be detected in written addition	5	NOG5IM17	1	
Position of error to be detected in written subtraction	5	NOG5IM16	1	
Reading/sorting of numbers when digit 0 is involved	5	NOG5IM9	1*	
	5	NOG5IM18	1*	
Sorting of numbers with different amount of digits	5	NOG5IM9	1*	
Position of missing summand out of 3	5	NOG5IM10	1	
Total			**34 (26*)**	**2(2*)**

*indicates significance ($p < .05$) of a factor.

In total, explanation of item instances appears to be relatively stable across grades at a surprisingly high level. Fewer parameters in general, and item models needing only one parameter in Grade 5, however, point to a more feasible modeling for higher grades. This could be explained twofold: First, by the vanishing impact of certain cognitive factors, for example, inversion or carry-over becoming non-significant in higher grades (see Table 2), probably due to being overlearned. Second, it appears that item modeling is noisier and more difficult in lower grades, likely due to factors not being/or not possible to be considered in the design nor being detectable in exploration (nor tackled in relevant cognitive literature).

Space & Shape. Compared to *NO*, fewer item instances generated for *SS* were fully explained (Figure 2, right). Only 52.5% of the 120 items administered in Grade 1 could be successfully modeled ($M = 78.12\%$, $SD = 33$). While three item models were fully explained, one model (SSG1IM8, consisting of 48 items) could not be grouped at all. Importantly, all factors used for grouping originated from the design (see Table 3).

In Grade 3, explanation rates remained low; only 51.28% of the 79 item instances could be explained ($M = 50\%$, $SD = 42.72$). Three item models consisted of individual items only, and one item model could be explained 100%. As in Grade 1, all factors used for grouping were based solely on the design. Notably, two of the item models relied on just a single factor for explanation.

Table 3. Overview on studied cognitive factors in *Space & Shape* to explain item difficulty of instances.

Cognitive factor	Grade	IM	design	explored
Number of Rows	1	SSG1IM1	1*	
	1	SSG1IM2	1*	
	3	SSG3IM1	1	
	3	SSG3IM2	1*	
Shape	1	SSG1IM3	1*	
	1	SSG1IM8	1*	
	3	SSG3IM9	1*	
	3	SSG3IM2	1*	
	3	SSG3IM1	1	
	3	SSG3IM3	1*	
	3	SSG3IM10	1*	
	5	SSG5IM11	1	
Number of Colors	1	SSG1IM4	1*	
Number of Spaces	1	SSG1IM4	1*	
	3	SSG3IM3	1*	
	5	SSG5IM15	1	
	5	SSG5IM16	1	
	5	SSG5IM9	1	
Pattern	1	SSG1IM4	1	
	1	SSG1IM5	1	
	3	SSG3IM5	1	
	5	SSG5IM15	1	
	5	SSG5IM16	1	
Continuous Design (Mirroring)	1	SSG1IM6	1*	
	1	SSG1IM7	1*	
	3	SSG3IM6	1*	
	3	SSG3IM7	1*	
Aligned symmetry Axis (Mirroring)	1	SSG1IM6	1*	
	1	SSG1IM7	1*	
	3	SSG3IM6	1*	
	3	SSG3IM7	1*	
Direction of Mirroring	1	SSG1IM6	1*	
	3	SSG3IM6	1*	
	3	SSG3IM7	1*	
Rotation	1	SSG1IM8	1*	
	3	SSG3IM10	1*	
Size of Area	3	SSG3IM9	1*	
	5	SSG5IM12	1*	
	5	SSG5IM13	1*	
	5	SSG5IM14	1	
	5	SSG5IM9	1	
Total			**41(28*)**	**0(0*)**

*indicates significance ($p < .05$) of a factor.

Grade 5 showed comparable results: Of the 54 item instances, 59.25% could be explained ($M = 66.67$, $SD = 30{,}86$). As with lower grades, all factors used for grouping were derived solely from the design. Two item models were fully explained, while one model required individual parameters for all items (SSG5IM17, 12 items). Additionally, grouped item instances of four models were explained using just one design-derived factor.

It seems that item modeling for *SS* was much less consistent and thus, in total, less successful compared to *NO*. Although the number of models with fully explained item instances was only slightly lower (6 for *SS* vs. 8 for *NO*), there were more models for which items couldn't be grouped at all (5 for *SS* vs. 1 for *NO*), partly with many item instances. The most problematic item models (SSG1IM8, SSG3IM10, SSG5IM17) longitudinally tapped the same subcompetency (analysis and representation of geometric shapes in plane and space; levels B1, B3, and B6). It therefore seems that measuring this construct seems to be noisier than expected or harder to cognitively model. The specific item type of these models (select one geometric form with differing characteristics) might be an alternative explanation.

Overall, cognitive modeling of the developed 48 item models was highly successful. Whereas item difficulty variance of 82.29% of 360 item instances within *NO* could be explained, within *SS*, this was true for 64.93% of the 252 item instances.

3.1.2. Identified Cognitive Factors

Within *NO*, out of the 20 cognitive grouping factors (see Table 2), most were known in relevant literature on number development and therefore already considered in the item design. Factors such as inversed ordering, number range, borrowing, or carry-over were among the strongest (stable) predictors of item difficulty. Exploration showed that the orientation of objects (horizontal vs. vertical) to compare and sort – a spatial aspect -, and whether the difference in a calculation was even or odd (both in Grade 1), were factors significantly contributing to item difficulty that were previously unknown. The number of significant factors was comparable across grades (8 for Grade 1, 6 for Grade 3, and 7 for Grade 5). However, Grade 5 featured five additional factors implemented in the design that turned out insignificant, with three of them significantly contributing in lower grades (inversion, number range, carry-over). This pattern suggests that the influence of these factors declines with age, consistent with developmental research showing that math fluency in elementary school increases with age and practice, while the trajectory and timing of automaticity vary by operation and task characteristics (e.g. Gliksman et al., 2022). Consequently, certain cognitive factors that substantially affect difficulty in lower grades may diminish in influence once skills become overlearned.

In the *SS* domain, 10 cognitive factors were incorporated into the item design, with 9 of them significantly influencing at least some item models. No additional cognitive factors were discovered during the exploration process (see Table 3 for studied cognitive effects within *SS*). The impact of these factors was comparable in Grades 1 and 3, but in Grade 5, only one factor significantly affected item difficulties. This pattern aligns with findings from *NO*, which also showed a diminishing influence of (known and manipulable) cognitive factors among older students. The overall lower number of factors, compared to *NO*, further reflects the relatively limited research in this mathematical domain. Additionally, the lower percentage of explained item instances suggests that other factors, not yet considered, are likely influencing item difficulty.

Taken together, results show that cognitive factors explain item difficulty in *NO* well, enabling reliable control in item design based on proto-theories. This effect is strongest in lower grades and diminishes as overlearned skills reduce the influence of these factors. In *SS*, however, modeling was less consistent, particularly for geometric shape analysis, indicating that additional or alternative factors are needed. Thus, while *NO* items can be effectively calibrated using established cognitive characteristics, *SS* requires further development of our cognitive models to identify additional factors for comparable precision. When applied to AIG, this means that reliable automated generation is currently more feasible for *NO* than for *SS*, where item difficulty remains harder to predict. Building on these results, we next examine to what extent contextual embeddings of problems further account for variability in item difficulty.

3.1.3. Explainability of Item Difficulty Variance by Contextual Factors

Within the *NO* domain, 6 out of 7[1] item models in Grade 1 (85.71%, see Figure 2) were affected by contextual factors, i.e. a variation of the context in which the mathematical problems were presented had a significant impact on item difficulty. This impact was comparable in Grade 3 (7 out of 8 item models showed an effect) but substantially dropped to only 2 affected item models in Grade 5. Table 4 shows that in most cases, it was one specific context that moderated item difficulties; in 3 item models, differences between contextualized and decontextualized (i.e. abstract) item instances were found. Similar to cognitive factors, more statistically significant effects were found in lower grades but not in

[1]In *Numbers & Operations*, Grade 1, item instances of one item model (NOG1IM7) were only presented in a decontextualized, abstract form due to an error in the booklet design.

Table 4. Overview on studied contextual factors in *Numbers & Operations* (left) and *Space & Shape* (right) to explain item difficulty of instances.

Contextual factor	Grade	IM	design	explored	Contextual factor	Grade	IM	design	explored
Decontextualization	1	NOG1IM1	1*		Decontextualization	1	SSG1IM3	1	
	1	NOG1IM2	1			1	SSG1IM6	1	
	1	NOG1IM3	1			1	SSG1IM7	1	
	3	NOG3IM10	1*			1	SSG1IM8	1*	
	3	NOG3IM14	1			3	SSG3IM3	1*	
	5	NOG5IM10	1			3	SSG3IM6	1	
	5	NOG5IM11	1*			3	SSG3IM7	1	
Gender stereotypical context	1	NOG1IM1	1*			3	SSG3IM10	1*	
	5	NOG5IM16	1			5	SSG5IM11	1*	
	5	NOG1IM17	1			5	SSG5IM12	1	
Specific context	1	NOG1IM4	1*			5	SSG5IM13	1	
	1	NOG1IM5	1*			5	SSG5IM14	1	
	1	NOG1IM6	1*			5	SSG5IM15	1	
	1	NOG1IM8	1*			5	SSG5IM16	1	
	3	NOG3IM4	1*			5	SSG5IM17	1*	
	3	NOG3IM8	1*		Specific context	1	SSG1IM1		1*
	3	NOG3IM9	1*			1	SSG1IM2		1*
	3	NOG3IM10	1*			1	SSG1IM4	1	
	3	NOG3IM11	1*			1	SSG1IM5		1*
	3	NOG3IM12	1*			3	SSG3IM1	1	
	3	NOG3IM13	1*			3	SSG3IM2	1*	
	5	NOG5IM9	1			3	SSG3IM5	1	
	5	NOG5IM10	1*			3	SSG3IM9	1*	
	5	NOG5IM11	1*			5	SSG5IM9	1	
	5	NOG5IM13	1						
	5	NOG5IM15	1						
Total			**26 (17*)**	**0(0*)**				**21(7*)**	**3(3*)**

*indicates significance ($p < .05$) of a factor.

Grade 5. For example, the impact of a gender stereotypical embedding was only significant in Grade 1. No additional contextual factors were identified beyond those implemented by design.

Item models of *SS* were impacted much less by contextual factors. Whereas in Grade 1, half of the item models were affected, this number dropped almost linearly to 25% of the models in Grade 5. Item difficulty differences were equally caused by decontextualized and specific embeddings. Taken together, results clearly show that the way a mathematical problem is presented clearly matters. This effect is more pronounced for younger students and diminishes over time to, as it seems, a baseline noise of 25% of models in Grade 5. Mostly, individual specific contexts caused differences with no clear rule or trend concerning the content. Thus, if different contexts are used for AIG, context needs to be considered, as it moderates the difficulty of classes of items based on cognitive characteristics.

3.2. RQ2 - Fairness of Item Models

3.2.1. Impact of Cognitive Radicals on Subgroup Differences

After identifying cognitive item characteristics that influenced item difficulty, we investigated whether these characteristics were associated with subgroup differences that could be interpreted as *Differential Radical Functioning* (DRF, see above). For instance, in item model *NOG1IM6*, students were asked to add dots to mushrooms until reaching a specific total. When analyzing item model validity, number range turned out to substantially influence item instance difficulty. Interestingly, girls found it easier than boys to complete the task when the target sum was 10, indicating a significant interaction between gender and this cognitive parameter, even when controlling for main effects and other interactions in a related GLMM. However, this effect did not occur when the target sum was 5, suggesting that the number range parameter exhibited DRF in this item model. A possible explanation

for this observation could be previously documented gender-specific strategy use in arithmetic problems (e.g. Fennema et al., 1998; Sievert et al., 2025): girls were found to rely more on concrete strategies, such as counting, or in later grades, written algorithms, which tend to be slower but more reliable. Boys, in contrast, more often use retrieval or mental computations (faster but potentially more error-prone non-algorithmic strategies or shortcut methods). Figure 3 shows the total number of item models that demonstrated one or more DRFs in relation to student background.

In the domain of *NO*, Math Interest (significant in 12 out of 24 item models), Math Self-Concept (11 out of 24), and Gender (11 out of 24) exhibited the most DRFs. A similar pattern was observed in the domain of *SS* where Gender (12 out of 24) and Math Interest (10 out of 24) had the highest number of DRFs. On average, each background variable influenced about one-third of the cognitive parameters. For *NO*, the number of DRFs slightly increased from Grade 1 to Grade 5, whereas the opposite trend was observed for *SS*, with a sharp drop in Grade 5, where only 10 DRFs were identified.

In summary, altering cognitive radicals when generating item instances can influence or uncover subgroup differences. Although this may not compromise test fairness, as cognitive aspects are integral to the construct (e.g. counting or performing addition with different number ranges), this should be done consciously; otherwise, such items could potentially be flagged for DIF.

3.2.2. Impact of Contextual Radicals on Subgroup Differences

In contrast to subgroup differences arising from cognitive radicals, those related to contextual radicals are more problematic and can compromise test fairness. Ideally, any increase in item difficulty due to contextual elements should affect all students uniformly, rather than disproportionately impacting students with specific background characteristics. Fortunately, our analysis of interactions between item models' contextual effects and student background characteristics revealed significantly fewer DRFs linked to contextual radicals compared to cognitive ones (see Figure 3).

For both mathematical domains, the largest number of contextual DRFs was associated with students' migration background, affecting eight item models each. Crucially, language background showed much less impact, suggesting that certain contexts may pose greater difficulty across different cultural backgrounds, regardless of language proficiency. This finding offers a new perspective on subgroup differences typically attributed to language proficiency, indicating that cultural context may

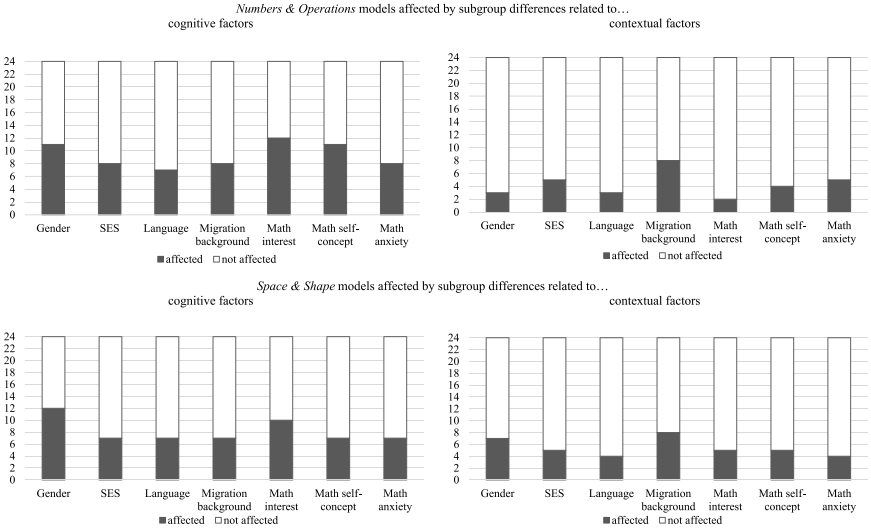


Figure 3. Number of item models with identified cognitive radicals (left) and contextual radicals (right) associated with significant subgroup differences.

play a larger role. Gender was identified as the second most frequent cause of contextual DRFs, though this was only observed in the *SS* domain.

Regarding grade-level differences, a notable increase in DRFs was observed in Grade 3 of the *SS* domain, which also featured a larger number of items that couldn't be grouped with others and therefore needing their own parameter. It remains unclear whether these items functioned differently due to related subgroup differences or if individual treatment led to these differences. Overall, an average of 4.28 affected models in *NO* and 5.42 models in *SS* indicates that subgroup differences tied to contextual elements are worth considering but do not present a major concern in AIG. Finally, a content review of contexts that exhibited DRF across subgroups revealed no recurring themes or other communalities. The observed differences appeared context-specific, tied to particular depictions within individual item models, and non-systematic across contexts, suggesting that such effects are hard to anticipate without indications for guidelines.

4. General Discussion

The use of automated item generation (AIG) in large-scale educational assessments (LSAs) holds promise for addressing the growing demand for cost-efficient and validated assessment items. However, despite the theoretical appeal of strong-theory AIG (e.g. Drasgow et al., 2006) for improving construct validity, empirical evaluations in applied settings remain scarce (Embretson & Kingston, 2018; Gierl et al., 2015). The present study addressed this gap by investigating the validity and fairness of 48 cognitive item models in the mathematical domains of *Numbers & Operations* and *Space & Shape*.

By treating template-based item models as proto-theories (e.g. Borsboom et al., 2021; Nettle, 2021), which offer greater flexibility than traditionally formulated cognitive models, and which allow for incorporating isolated effects or observations from practitioners, we explored to what extent AIG can be theoretically informed. Regarding RQ1 (validity of item models), our results indicate that the difficulty of AIG-generated items can be explained to a substantial extent by their predefined cognitive and contextual features, which demonstrated the high effectiveness of our approach. A substantial proportion of items (82.29% in *NO*, 64.39% in SS) was explained by cognitive and contextual factors, most of which were already embedded in the models used for generating them – an indication that overfitting did not occur, strengthening the validity of the proto-theory approach. Given the lack of comprehensive cognitive theories or their mismatch in granularity for many LSA-assessed competencies (Leighton & Gierl, 2011), these findings support a more pragmatic, theory-informed AIG approach that balances empirical observations and practitioner insights. For the ÉpStan, this study provided an empirical inventory of radicals and incidentals for *NO* and *SS*, offering a practical foundation for future item development and enhancing construct validity. The next essential step is a confirmatory study to test the effects' stability in newly generated items with identical features.

Another key contribution of this study is its examination of the impact and fairness of using different contexts when generating language-reduced, image-based items (RQ2 – fairness of item models). Our results show that contextual embeddings can influence the interpretive fairness of AIG-based items, with effects varying by student subgroup. In both explored mathematical domains, we observed that contextual influences on item difficulty decreased with age, indicating a shift toward more stable problem representations and improved transfer skills. These findings support developmental theories of mathematical cognition, suggesting that older students rely less on surface-level features and more on abstract problem-solving strategies (e.g. Chen, 1999). While minimizing linguistic complexity is one way to fair assessments, our study revealed that contextual variations still introduce sociocultural biases, particularly for students from migration backgrounds. These findings confirm previous research on sociocultural knowledge differences among immigrant students (Martin et al., 2012; Sam et al., 2022), emphasizing explicit consideration of cultural backgrounds when generating diverse problem contexts.

However, our review of contexts that exhibited DRF revealed no recurring themes, underscoring the need for systematic future research on how contextual embeddings contribute to subgroup differences.

Importantly, cognitive radicals that contributed to subgroup differences could provide valuable insights for math instruction, supporting the reciprocal relationship between assessment and teaching (De Lange, 2007). In sum, DRF analyses provide item-level evidence that can inform the interpretive fairness of AIG-based items by distinguishing construct-relevant subgroup differences from potential bias.

Despite these contributions, several limitations must be acknowledged. Due to the relatively small item pools per model, we were unable to perform detailed analyses of within- and between-item family variation, as seen in prior studies (Embretson & Kingston, 2018; Sinharay & Johnson, 2008). This limits direct comparability with earlier findings but also presents an opportunity for future research, as the developed item models could be used for larger-scale administration in future LSAs. Another limitation lies in the selection of contextual embeddings – while they were typical of Luxembourg's ÉpStan, their design was not systematically standardized, making further refinement necessary. However, this was due to limited available research on visual item contexts and points to an open area for future psychometric research that includes, for example, cognitive load theory.

In light of recent developments in generative AI, template-based AIG might seem less innovative at first glance. However, most current commercial AI-driven solutions still focus on text-based multiple-choice problems, while language-reduced, image-based items remain largely underexplored (Circi et al., 2023; Harris et al., 2024). Items using artificially generated images would still require careful human review before implementation and raise additional legal challenges, particularly in the European framework (European Parliament, 2024), regarding copyright and the black box nature of current large language models (LLMs). Rather than replacing theory-informed item modeling, the most promising way forward lies in combining both strands: AI can provide efficiency and scale, while cognitive models ensure transparency, validity, and fairness, an approach increasingly discussed under the labels of augmented intelligence or human-centered AI (Circi et al., 2023; Harris et al., 2024; A. A. von Davier et al., 2024; Yang et al., 2023).

To conclude, this study provides empirical evidence supporting the feasibility of strong-theory AIG in applied settings. While further validation and cross-study replication are needed, our findings mark a promising step toward a validated, template-based AIG approach for elementary mathematics education. Future research should focus on confirmatory studies, cross-national validation, and the integration of AI-driven enhancements while maintaining rigorous psychometric control.

Acknowledgments

We thank Michel Roeder for his contributions to the development of cognitive item models and Cécile Braun for her careful proofreading of the manuscript.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This study was funded by the Fonds National de la Recherche Luxembourg [grant FAIR-ITEMS (C19/SC/13650128)].

Research Data Policy and Data Availability Statements

The datasets generated during and/or analyzed during the current study are available from the first author on reasonable request.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. AERA.

Arendasy, M., & Sommer, M. (2007). Using psychometric technology in educational assessment: The case of a schema-based isomorphic approach to the automatic generation of quantitative reasoning items. *Learning and Individual Differences*, *17*(4), 366–383. https://doi.org/10.1016/j.lindif.2007.03.005

Attali, Y. (2018). Automatic item generation unleashed: An evaluation of a large-scale deployment of item models. In *Artificial Intelligence in Education: 19th International Conference AIED 2018* (pp. 17–29). London, UK: Springer International Publishing.

Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance? *Computers and Education*, *83*, 57–63. https://doi.org/10.1016/j.compedu.2014.12.012

Attali, Y., & Arieli-Attali, M. (2019). Validating classifications from learning progressions: Framework and implementation. *ETS Research Report Series*, *2019*(1), 1–20. https://doi.org/10.1002/ets2.12253

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blum, W., & Leiss, D. (2007). How do students and teachers deal with modelling problems. In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling: Education, engineering and economics-ICTMA* (Vol. 12, pp. 222–231). Horwood: Chichester.

Borsboom, D., H. L. Van Der Maas, J. Dalege, R. A. Kievit, & B. D. Haig. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16* (4), 756–766.

Chen, Z. (1999). Schema induction in children's analogical problem solving. *Journal of Educational Psychology*, *91*(4), 703. https://doi.org/10.1037/0022-0663.91.4.703

Cho, S. J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika*, *79*(1), 84–104. https://doi.org/10.1007/s11336-013-9360-2

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. In *Frontiers in education* (Vol. 8, p. 858273). https://doi.org/10.3389/feduc.2023.858273

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, *34*(5), 348–364. https://doi.org/10.1177/0146621609349801

Davis-Kean, P. E., Domina, T., Kuhfeld, M., Ellis, A., & Gershoff, E. T. (2022). It matters how you start: Early numeracy mastery predicts high school math course-taking and college attendance. *Infant and Child Development*, *31*(2), e2281. https://doi.org/10.1002/icd.2281

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28. https://doi.org/10.18637/jss.v039.i12

De Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 1111-1142). Charlotte, NC: Information Age Publishing.

Dewolf, T., Van Dooren, W., & Verschaffel, L. (2017). Can visual aids in representational illustrations help pupils to solve mathematical word problems more realistically? *European Journal of Psychology of Education*, *32*(3), 335–351. https://doi.org/10.1007/s10212-016-0308-7

Drasgow, F., Luecht, R., & Bennett, R. (2006). Technology and testing. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–516). Westport, CT: Praeger Publishers.

Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement*, *55*(1), 112–131. https://doi.org/10.1111/jedm.12166

European Parliament. (2024). Artificial intelligence act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.Pdf

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher*, *27*(5), 6–11. https://doi.org/10.3102/0013189X027005006

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic item generation*. Routledge.

Gierl, M. J., & Lai, H. (2016). A process for reviewing and evaluating generated test items. *Educational Measurement Issues & Practice*, *35*(4), 6–20. https://doi.org/10.1111/emip.12129

Gierl, M. J., Lai, H., Hogan, J. B., & Matovinovic, D. (2015). A method for generating educational test items that are aligned to the Common Core State Standards. *Journal of Applied Testing Technology*, *16*(1), 1–18. http://www.jattjournal.net/index.php/atp/article/view/80234.

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning, and Assessment*, *7*(2),1–51.
Girardelli, I. (2017). *Der Einfluss bildlicher Kontextualisierung auf die Schwierigkeit mathematischer Testaufgaben*. [The impact of visual contextualisation on the difficulty of mathematical tasks.] [Unpublished Master thesis]. University of Luxembourg.
Gliksman, Y., Berebbi, S., & Henik, A. (2022). Math fluency during primary school. *Brain Sciences*, *12*(3), 371. https://doi.org/10.3390/brainsci12030371
Greisen, M., Georges, C., Hornung, C., Sonnleitner, P., & Schiltz, C. (2021). Learning mathematics with shackles: How lower reading comprehension in the language of mathematics instruction accounts for lower mathematics achievement in speakers of different home languages. *Acta Psychologica*, *221*, 103456. https://doi.org/10.1016/j.actpsy.2021.103456
Harris, D. J., Welch, C. J., & Dunbar, S. B. (2024). In the beginning, there was an item . . .. *Educational Measurement Issues & Practice*, *43*(4), 40–45. https://doi.org/10.1111/emip.12647
Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.
Hoogland, K., Pepin, B., de Koning, J., Bakker, A., & Gravemeijer, K. (2018). Word problems versus image-rich problems: An analysis of effects of task characteristics on students' performance on contextual mathematics problems. *Research in Mathematics Education*, *20*(1), 37–52. https://doi.org/10.1080/14794802.2017.1413414
Irvine, S. H. (2002). The foundations of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3–34). Erlbaum.
Irvine, S., & Kyllonen, P. (2002). *Generating items for cognitive tests: Theory and practice*. Lawrence Erlbaum.
Kellogg, M., Rauch, S., Leathers, R., Simpson, M. A., Lines, D., Bickel, L., & Elmore, J. (2015, April). Construction of a dynamic item generator for K-12 mathematics. In *National Council on Measurement in Education Conference*. Chicago, IL: Na-tional Council on Measurement in Education Conference.
Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement Issues & Practice*, *38*(1), 48–53. https://doi.org/10.1111/emip.12237
Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge University Press.
Leiss, D., Ehmke, T., & Heine, L. (2024). Reality-based tasks for competency-based education: The need for an integrated analysis of subject-specific, linguistic, and contextual task features. *Learning and Individual Differences*, *114*, 102518. https://doi.org/10.1016/j.lindif.2024.102518
Martin, A. J., Liem, G. A. D., Mok, M. M. C., & Xu, J. (2012). Problem solving and immigrant student mathematics and science achievement: Multination findings from PISA. *Journal of Educational Psychology*, *104*(4), 1054–1073. https://doi.org/10.1037/a0029152
MENFP. (2011). *Plan d ' études. École fondamentale* MENFP.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825
National Center for Education Statistics. (2022). *The nation's report card: Mathematics 2022*. NCES 2023-001. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/
Nettle, D. (2021, May 16). Theories and models are not the only fruit. *Medium*. https://leotiokhin.medium.com/theories-and-models-are-not-the-only-fruit-a05c7cf188f6
Nohara, D. (2001). A comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA). Working paper no. 2001-07. National Center for Educational Statistics.
OECD. (2012). *Untapped skills: Realising the potential of immigrant students*.
OECD. (2024). *Education at a glance, 2024: OECD indicators*.
R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org
Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test Analysis Modules*. R package version 4.2-21. https://CRAN.R-project.org/package=TAM
Saalbach, H., Gunzenhauser, C., Kempert, S., & Karbach, J. (2016). *Der Einfluss von mehrsprachigkeit auf mathematische fähigkeiten bei grundschulkindern mit niedrigem sozioökonomischen status*. Frühe Bildung.
Sam, D. L., Vedder, P., Ward, C., & Horenczyk, G. (2022). Psychological and sociocultural adaptation of immigrant youth. In J. W. Berry, J. Phinney, & D. Sam (Eds), *Immigrant youth in cultural transition* (pp. 119–143). Routledge.
Sells, L. (1973). High school mathematics as the critical filter in the job market. In R. T. Thomas (Ed.), *Developing opportunities for minorities in graduate education* (pp. 37–39). University of California Press.
Sievert, H., Hickendorff, M., van den Ham, A. K., & Heinze, A. (2025). Children's arithmetic strategy use and strategy change from grade 3 to grade 4. *International Journal of Science & Mathematics Education*, 1–22. https://doi.org/10.1007/s10763-025-10578-3
Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: A meta-analytic study. *School Psychology Quarterly*, *32*(4), 435. https://doi.org/10.1037/spq0000197

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Lawrence Erlbaum.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing*, *8*(3), 209–236. https://doi.org/10.1080/15305050802262019

Sonnleitner, P., Krämer, C., Gamo, S., Reichert, M., Muller, C., Keller, U., & Ugen, S. (2018). *Schülerkompetenzen im Längsschnitt - Die Entwicklung von Deutsch-Leseverstehen und Mathematik in Luxemburg zwischen der 3. und 9. Klasse*. LUCET, Universität Luxemburg, SCRIPT.

Van Dooren, W., & Inglis, M. (2015). Inhibitory control in mathematical thinking, learning and problem solving: A survey. *ZDM*, *47*(5), 713–721. https://doi.org/10.1007/s11858-015-0715-2

von Davier, A. A., Runge, A., Park, Y., Attali, Y., Church, J., & LaFlair, G. (2024). The item factory. In Jiao Hong & Robert W. Lissitz (Eds.), *Machine learning, natural language processing, and psychometrics* (pp. 1-26). Information Age Publishing.

von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *Timss, 2023 technical report (methods and procedures)*. TIMSS & PIRLS International Study Center.

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932. https://doi.org/10.1037/a0031882

Wilson, J., Morrison, K., & Embretson, S. E. (2014). *Automatic item generator for mathematical achievement items: MathGen3.0* (Technical report IES1005A-2014). Cognitive Measurement Laboratory, Georgia Institute of Technology.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers. Theory into practice*. Springer.

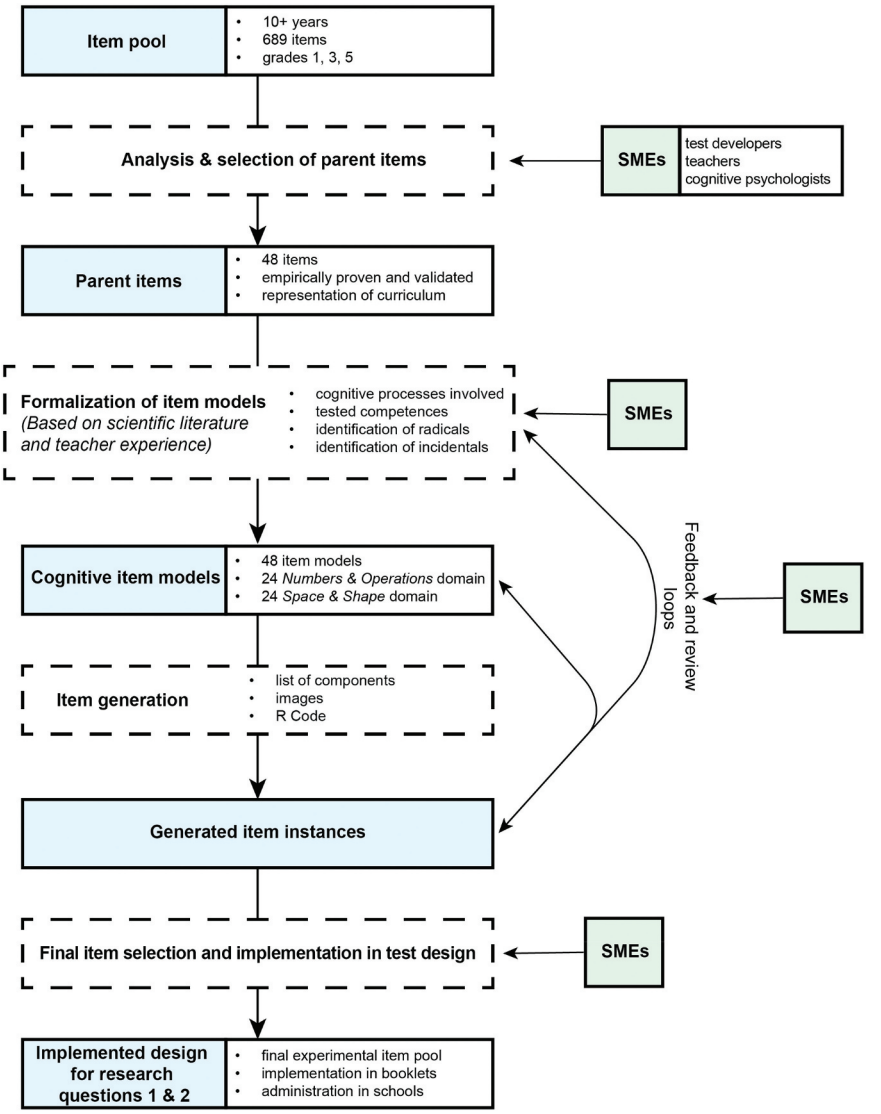
Yang, S. J. H., Ogata, H., & Matsui, T. (2023). Guest editorial: Human-centered AI in education: Augment human intelligence with machine intelligence. *Educational Technology & Society*, *26*(1), 95–98.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Annexes

Annex 1. Schematic flowchart of cognitive item model development and implementation in R building on Gierl et a (2015)The process runs from parent item identification to generation of item instances and integration into the test design SMEs= subject-matter experts



Annex 2. Overview of item models and generated item instances for the domain Numbers & Operations in study 1,

		Tested subcompetency (item models with identical subcompetency differ in their task model or response format)	Tested subcompetency levels*	Item instances	N° of parameters	N° of outliers
Numbers & Operations (ÉpStan 2021)						
Grade 1	NOG1IM1	The student is able to count and compare sets of objects with up to 10 elements.	A1	12	9	0
	NOG1IM2	The student can associate quantifiable objects with up to 5 numbers and sort them accordingly.	B1	12	6	0
	NOG1IM3	The student can associate quantifiable objects with up to 5 numbers and sort them accordingly.	B1	12	2	0
	NOG1IM4	The student can match numbers from 0 to 20 with their symbols and arrange them in the correct order.	A2-A3	24	6	2
	NOG1IM5	The student can mentally calculate additions and subtractions within the range of 0 to 20.	B2	24	5	13
	NOG1IM6	The student can match numbers from 0 to 20 with their symbols and arrange them in the correct order.	A2	12	4	0
	NOG1IM7	The student can mentally calculate additions and subtractions within the range of 0 to 100.	B3	48	20	0
	NOG1IM8	The student can mentally calculate additions and subtractions within the range of 0 to 20.	B2	12	4	0
Grade 3	NOG3IM4	The student is able to read and write numbers from 0 to 100, as well as compare and sort them.	A3	18	3	0
	NOG3IM8	The student can mentally calculate additions and subtractions within the range of 0 to 100.	B3	12	4	1
	NOG3IM9	The student can read and write natural numbers from 0 to 1,000, compare and sort them.	A4	12	3	1
	NOG3IM10	The student can perform additions and subtractions with up to 4 significant digits within the range of 0 to 1,000.	B4	12	6	0
	NOG3IM11	The student can perform additions and subtractions with up to 4 significant digits within the range of 0 to 1,000.	B4	12	6	0
	NOG3IM12	The student can perform additions and subtractions with up to 4 significant digits within the range of 0 to 1,000.	B4	24	0	24
	NOG3IM13	The student can perform additions and subtractions with up to 4 significant digits within the range of 0 to 1,000.	B4	12	5	2
	NOG3IM14	The student differentiates between even and odd numbers.	C2	12	2	2
Grade 5	NOG5IM9	The student can read and write natural numbers from 0 to 1,000,000, compare, and present them in a table.	A4–6	12	3	1
	NOG5IM10	The student can perform additions and subtractions with up to 4 significant digits within the range of 0 to 1,000.	B4	18	6	1
	NOG5IM11	The student can perform additions, subtractions, and multiplications within the range of 0 to 100,000.	B5–6	18	7	0
	NOG5IM13	The student can perform additions, subtractions, and multiplications within the range of 0 to 100,000.	B5	12	3	0
	NOG5IM15	The student can present data in a well-organized table, draw clear conclusions, and explain them.	D6	12	6	0
	NOG5IM16	The student can perform multiplications and divisions with up to 5 significant digits in writing.	B6	6	1	0
	NOG5IM17	The student can perform multiplications and divisions with up to 5 significant digits in writing.	B6	6	1	0
	NOG5IM18	The student can identify adjacent numbers and insert natural numbers from 0 to 1,000,000, as well as extract the values of units, tens, hundreds, etc., from a given number.	A6	6	5	0
Total	24			360	118	47

*Tested subcompetency level according to the Luxembourgish elementary school mathematics curriculum (cf. MENFP, 2011).

Annex 3. Overview of item models and generated item instances for the domain Space & Shape in study 2,

		Tested subcompetency (item models with identical subcompetency differ in their task model or response format)	Tested subcompetency levels	Item instances	N° of parameters	N° of outliers
Space & Shape (ÉpStan 2022)						
Grade 1	SSG1IM1	Starting with simple shapes, the student can create various figures that have the same area.	D1	12	2	0
	SSG1IM2	In the context of basic counting, the student can distinguish between area and perimeter.	D3	12	4	1
	SSG1IM3	The student can group different simple shapes represented on squared paper that have the same area.	D2	12	3	1
	SSG1IM4	The student can reproduce and extend basic geometric patterns and structures.	C1	12	2	1
	SSG1IM5	The student can reproduce and extend basic geometric patterns and structures.	C1	6	2	0
	SSG1IM6	The student can complete basic geometric shapes using axial symmetry.	C3	12	6	1
	SSG1IM7	The student can complete basic geometric shapes using axial symmetry.	C3	6	3	0
	SSG1IM8	The student can identify, compare, and group quadrilateral, triangular, and circular shapes.	B1	48	48	0
Grade 3	SSG3IM9	The student can use counting to determine the area and perimeter of basic shapes (square, rectangle).	D4	6	6	0
	SSG3IM2	In the context of basic counting, the student can distinguish between area and perimeter.	D3	6	6	0
	SSG3IM1	The student can group various simple shapes represented on squared paper that have the same area.	D2	6	1	1
	SSG3IM3	In the context of basic counting, the student can distinguish between area and perimeter.	D3	12	3	1
	SSG3IM10	The student can use the correct terms to name flat shapes and basic solids.	B3	18	18	0
	SSG3IM5	The student can extend complex geometric patterns and create his own unique patterns.	C4	6	1	1
	SSG3IM6	The student can extend complex geometric patterns and create his own unique patterns.	C4	12	6	0
	SSG3IM7	The student can extend complex geometric patterns and create his own unique patterns.	C4	12	6	1
Grade 5	SSG5IM11	The student can compare or determine the area of any right-angled shape by breaking it down into uniform units.	D6	6	2	0
	SSG5IM12	The student can compare or determine the area of any right-angled shape by breaking it down into uniform units.	D6	6	3	0
	SSG5IM13	The student can measure the perimeter of a square and a rectangle by filling them with uniform units.	D5	6	3	1
	SSG5IM14	The student can determine the area and perimeter of basic shapes shown on squared paper by counting.	D4	6	1	1
	SSG5IM15	The student can extend complex geometric patterns and create his own unique patterns.	C4	6	1	1
	SSG5IM16	The student can extend complex geometric patterns and create his own unique patterns.	C4	6	1	1
	SSG5IM9	The student can determine the area and perimeter of basic shapes shown on squared paper by counting.	D4	6	1	2
	SSG5IM17	The student can draw geometric shapes based on the properties of straight lines and line segments.	B6	12	12	0
Total	24			252	141	14

*Tested subcompetency level according to the Luxembourgish elementary school mathematics curriculum (cf. MENFP, 2011).