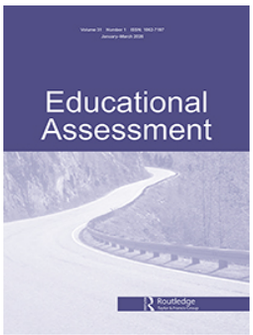


Routledge
Taylor & Francis Group

Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: www.tandfonline.com/journals/heda20

Mapping First Grader's Numerical Development at Scale: Leveraging Cognitive Models in a Large-Scale Educational Assessment

Philipp Sonnleitner, Pamela Inostroza Fernández & Caroline Hornung

To cite this article: Philipp Sonnleitner, Pamela Inostroza Fernández & Caroline Hornung (26 Feb 2026): Mapping First Grader's Numerical Development at Scale: Leveraging Cognitive Models in a Large-Scale Educational Assessment, Educational Assessment, DOI: 10.1080/10627197.2026.2633512

To link to this article: https://doi.org/10.1080/10627197.2026.2633512

Published online: 26 Feb 2026.

EDUCATIONAL ASSESSMENT
https://doi.org/10.1080/10627197.2026.2633512

Routledge
Taylor & Francis Group

Mapping First Grader's Numerical Development at Scale: Leveraging Cognitive Models in a Large-Scale Educational Assessment

Philipp Sonnleitner[a], Pamela Inostroza Fernández[b], and Caroline Hornung[a]

[a]Luxembourg Centre for Educational Testing, University of Luxembourg; [b]Innovation Department, Universidad de los Andes

ABSTRACT

Large-scale assessments (LSAs) primarily support system-level monitoring, but their instructional diagnostic potential remains underused. Cognitive Diagnostic Models (CDMs) offer a promising avenue, though their application in LSAs poses theoretical and practical challenges. This study explores whether cognitive models used in item generation can directly inform Q-matrix construction for CDM analyses, enhancing diagnostic value in early numeracy assessment. Using data from Luxembourg's school monitoring program ($N = 2{,}704$), we analyzed four cognitive attributes (counting, addition < 10, decomposition, addition > 10) using developmental frameworks. We compared a Single-Attribute Hierarchical Model, assuming linear progression, with a Multiple-Attribute Hierarchical Model, allowing skill interactions. Both hierarchical models reproduced expected developmental progressions, with decomposition emerging as a key threshold skill and socio-economic status showing the largest subgroup differences. Subgroup analyses revealed a smaller-than-expected impact of migration background, while math anxiety peaked at intermediate skill levels. Embedding cognitive models in LSAs can bridge system-level monitoring and instructional support.

Introduction

Mathematical skills are foundational to modern societies, particularly in knowledge-based economies (Hoogland & Tout, 2018; OECD, 2024). They strongly predict key life outcomes, shaping academic success, career opportunities, and broader cognitive competencies (Davis-Kean et al., 2022; Sells, 1973). The increasing integration of digital technologies and the growing focus on data analytics across all sectors further amplify the importance of mathematics education, as it serves as the entry point to STEM-related fields that drive innovation and technological progress (National Council of Teachers of Mathematics, 2017). Consequently, mathematics remains a core focus of large-scale educational assessments (LSAs), which evaluate and monitor national and international education programs. Prominent LSAs, such as TIMSS, PISA, and the U.S. NAEP, inform policymakers and educational leaders about student achievement and curriculum effectiveness (National Center for Education Statistics, 2022; OECD, 2024; von Davier et al., 2024).

Despite their well-established role in educational monitoring, LSAs face significant limitations at the classroom level. The National Research Council (Pellegrino, Chudowsky & Glaser, 2001) identified three primary purposes of LSAs: (a) evaluating programs, (b) measuring whether students meet

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

This article was originally published with errors, which have now been corrected in the online version. Please see Correction (https://doi.org/10.1080/10627197.2026.2643024)

learning goals, and (c) assisting learning. While LSAs effectively serve policymakers and system-level decision-makers, they provide limited utility for teachers and students due to the often lower reliability of subscores and high error variance at the individual level. This misalignment of stakeholder needs has led to growing frustration among educators – those directly responsible for fostering mathematical skills – who feel that LSA feedback (often a single score for an entire domain) is too disconnected from instruction, failing to justify the substantial time investment required for test administration (Bennett, 1999; Huff & Goodman, 2007; Sessoms & Henson, 2018). Mathematics assessments are no exception (De Lange, 2007; Holliday & Holliday, 2003), with many questioning whether the insights gained from LSAs translate into meaningful, actionable feedback for teaching and learning.

A less discussed issue is the underuse of assessment expertise. While LSAs are developed collaboratively by psychometricians, subject-matter experts, and teachers to ensure validity (Wu et al., 2016), this contrasts with teachers' assessment literacy skills. DeLuca et al. (2020) found that teachers struggle to apply new assessment approaches in practice. This gap hinders the translation of LSA insights into classroom use. Given the considerable investment LSAs require, the question arises: Can population-based LSAs be designed to better support instruction without compromising their monitoring role?

Cognitive diagnostic models as promising additions to large-scale assessments

To provide more instructionally relevant feedback and better meet educators' needs, CDMs have been increasingly applied and proposed as a promising complement to traditional LSA methodologies (De La Torre, 2009; Leighton & Gierl, 2007). Unlike standard psychometric models, which primarily yield unidimensional summative scores, CDMs aim to capture students' mastery of specific cognitive attributes – underlying processes, skills, or knowledge structures required to solve an item. The foundation of CDMs is the so-called Q-matrix, linking item features to underlying attributes. The Q-matrix is traditionally defined by subject matter experts (SMEs) and specifies which cognitive skills are required for each item (by assigning a "1" if an attribute is measured by an item, or a "0" if not), forming the basis for estimating mastery profiles via CDMs (Leighton & Gierl, 2007). Despite their theoretical appeal, CDMs remain underutilized in applied educational assessments. Sessoms and Henson (2018), in their comprehensive review of the field, found that only half of published CDM papers (36 in total) focused on empirical applications using operational assessment data rather than simulations. Notably, many available studies focus on mathematics, including attempts to reanalyze LSA data using CDMs – such as for the TIMSS for 4th graders (Yamaguchi & Okada, 2018), the TIMSS for 8th graders (Choi et al., 2015), and PISA (e.g. Wu et al., 2020; Zhang et al., 2025).

This so-called retrofitting approach – common in LSA contexts – seeks to extract diagnostic information by assigning cognitive attributes post hoc. However, it faces several key limitations (Gierl & Cui, 2008; Liu et al., 2018). First, LSAs are typically designed to yield unidimensional scales for system monitoring, while CDMs require a multidimensional framework. As a result, items capturing multiple attributes are often excluded during test construction, limiting CDM applicability. Second, CDMs are typically applied in a confirmatory manner when validating hypothesized attribute structures. Hence, their validity depends on theoretically grounded Q-matrices established before item development. Retrofitting introduces a risk of misalignment with the item's original design. Third, because item development, cognitive modeling, and feedback mechanisms are not integrated, retrofitted CDMs rarely yield meaningful instructional insights. Consequently, students are often classified in extreme latent classes – either mastering all or none of the attributes – leaving little room for nuanced differentiation (e.g., Choi et al., 2015; Haghayeghi et al., 2025; Liu et al., 2018; Wu et al., 2020). This lack of granularity limits the instructional value of CDMs in LSA settings.

Despite these challenges, some studies have successfully integrated CDMs into diagnostic mathematics assessments designed with strong cognitive models from the outset. For example, Gierl et al. (2010) embedded CDMs into an operational assessment framework for grades 3 and 6, incorporating 26 cognitive models and 178 skills. Similarly, Kunina-Habenicht et al. (2017) applied CDMs to a diagnostic arithmetic assessment in 4th grade, focusing on the mathematical domain

Numbers & Operations. Examining the incremental validity of CDM scores compared to traditional unidimensional large-scale assessments, they found that while multidimensional CDM-based proficiency profiles were moderately correlated with both external criteria (teacher-assigned mathematics grades and a national standards-based mathematics test), they added only negligible incremental variance ($\Delta R^2 = 0.3\%–1.6\%$) beyond the unidimensional IRT score. This finding underscores that CDMs, although not necessarily designed to predict grades, primarily offer diagnostic value by distinguishing skill profiles rather than by enhancing predictive power for external outcomes.

Taken together, existing research suggests that CDMs are most effective when cognitive models are integrated from the start. However, such applications remain rare, and much of CDMs' potential in LSAs is still underexplored. As Gierl and Cui (2008) emphasized, successful CDM implementation requires fundamental shifts in test development practices, including early collaboration between content experts, cognitive psychologists, and psychometricians. Moreover, the granularity of existing cognitive models does not always align with the broader theoretical levels assessed in LSAs (e.g., Leighton & Gierl, 2007), making direct implementation of such models not always feasible. Likewise, as Tatsuoka et al. (2016) noted, defining and refining cognitive attributes is an iterative process that requires the integration of cognitive theory, item design, and psychometric modeling – a coordination that is rarely achieved in large-scale assessment contexts.

Why assessing early numerical development in LSAs matters

Recent studies confirm the feasibility of using CDMs for early numeracy assessment, supporting their potential in LSAs. Li et al. (2020) assessed kindergartners' mathematical problem-solving using 11 cognitive attributes, while Haghayeghi et al. (2025) applied CDMs to a first-grade checklist with nine attributes. Although the latter retrofitted the attribute mapping, both studies identified a clear skill hierarchy and demonstrated that CDMs can generate diagnostic insights that directly inform instruction. Further empirical evidence supports the instructional value of such diagnostic feedback, Tatsuoka (1992) showed that providing students with feedback based on their cognitive profiles led to measurable learning gains in fraction understanding among middle-school students.

Integrating CDMs into LSAs at this early age offers several advantages. First, early numeracy skills are strong predictors of later academic success, influencing mathematics achievement in elementary (Braeuning et al., 2021; Hornung et al., 2014; Jordan et al., 2009) and secondary education (Watts et al., 2014), as well as broader educational attainment (Claessens et al., 2009). Children struggling with mathematics tend to experience persistent difficulties and slower learning growth (Nelson & Powell, 2018), highlighting the importance of early identification and intervention. Second, early numeracy development is well-documented through strong cognitive models, making it particularly well-suited for CDMs. Krajewski and Schneider's (2009), extended by Roesch and Moeller (2015), describes three progressive levels: (1) Children acquire basic numerical skills, distinguishing quantities and reciting number sequences without necessarily grasping their numerical meaning. (2) They recognize that numbers describe set sizes, they acquire the cardinal number concept, use counting for enumerating and begin applying counting routines. (3) They develop an initial understanding of numerical relationships, and master decomposition (i.e. they understand that numbers can be broken down and recombined (e.g., 10 = 5 + 5 or 2 + 8), forming the basis for arithmetic operations). These stepwise, largely unidimensional learning trajectories align well with CDMs and may help overcome the multidimensionality issues encountered when applying CDMs to higher mathematics domains. Note that earlier studies applying CDMs with young children confirmed similar hierarchies (Haghayeghi et al., 2025; Li et al., 2020), suggesting that these developmental patterns are robust across contexts and making the use of this model particularly promising for CDM applications. Finally, early mathematics instruction is highly structured, closely following these developmental stages (e.g., the Luxembourgish Plan d´études, MENFP, 2011). Yet the instructional use of LSA feedback remains limited. Not only due to teachers' assessment literacy, but also because LSAs

typically provide only aggregated scores that lack diagnostic insight. Strengthening teachers' understanding of early mathematical development (Ban et al., 2024) and embedding CDMs in LSAs may jointly enhance the interpretability and classroom relevance of assessment feedback.

While CDMs have been successfully applied in mathematics research, most prior studies, particularly those in early numeracy, were conducted in small- to medium-scale research contexts (e.g., Haghayeghi et al., 2025; Li et al., 2020). LSA implementations remain rare and have mainly focused on upper-grade international programs such as TIMSS and PISA (e.g., De La Torre, 2011; Wu et al., 2020; Zhang et al., 2025). In line with the National Research Council (Pellegrino, Chudowsky & Glaser, 2001), LSAs are population-based assessments designed to monitor system-level educational outcomes rather than individual diagnostics. Embedding CDMs within such frameworks, such as the Luxembourg *Épreuves Standardisées* (*ÉpStan)*, offers a novel opportunity to link large-scale performance data to theoretically grounded cognitive models.

Diagnostic large-scale assessments such as the Dynamic Learning Maps and the Early Grade Mathematics Assessment illustrate that large-scale systems can generate instructionally useful feedback (cf. Clark & Karvonen, 2020; Platas et al., 2014). However, these programs target specific populations or contexts (e.g., students with cognitive disabilities or in developing countries) and are not designed for universal school monitoring. The present study extends this diagnostic perspective to a national cohort at the very beginning of formal schooling (Grade 1).

In sum, given their predictive value, strong theoretical grounding, and structured instruction, early numeracy skills are ideally suited for integrating CDMs into LSAs. Children who struggle with foundational concepts – such as linking numbers to quantities or understanding number relationships – are at greater risk of long-term math difficulties (e.g., Mazzocco & Thompson, 2005; M. Von Aster et al., 2007). As a result, these children may continue to use inefficient strategies such as positional finger counting in higher grades (Gersten & Chard, 1999) and hinder their overall mathematical development. Early identification and support can help prevent these challenges, maximizing instructional value while maintaining LSAs' system-level monitoring role.

Aims and scope of the study

Given the importance of early numeracy development and the diagnostic potential of CDMs, this study explores the feasibility and added value of applying CDMs within a large-scale assessment (LSA) context at the beginning of elementary school when interventions are promising. Specifically, we examine whether cognitive models used for item generation can serve as valid Q-matrices for CDM analyses and whether this approach yields meaningful insights at the cohort and group level, and allows for instructional feedback at the individual level.

The Luxembourg national school monitoring program, *Épreuves Standardisées* (*ÉpStan*), provides an interesting setting for this investigation due to its highly diverse and multilingual student population, where 43.5% of students hold a non-Luxembourgish nationality and 68.4% speak a different native language than the instruction language (Luxembourgish or German) (MENJE & SCRIPT, 2024). Family socio-economic status (SES) is a well-established predictor of early academic achievement, influencing children's exposure to mathematical activities and overall learning opportunities at home (Bradley & Corwyn, 2002; Davis-Kean et al., 2021; Jordan et al., 2009; Muñez et al., 2021). Migration status and home language background likewise affect early numeracy or numerical problem-solving by shaping proficiency in the instructional language (Eisenwort et al., 2018; Greisen et al., 2021; Sonnleitner et al., 2018). Moreover, gender-related strategy use has been observed even in basic arithmetic, with boys tending to employ retrieval or mental-calculation strategies and girls more often relying on counting approaches (Fennema et al., 1998; Sievert et al., 2025). Finally, motivational factors, such as interest in mathematics, or math anxiety, were identified to also influence math performance (e.g. Suárez-Pellicioni et al., 2015). By moving beyond a single unidimensional score, CDMs could offer a more detailed perspective on students' numerical development, helping to explain

performance differences between learner groups and inform instructional decisions. This study addresses two key research questions:

Can cognitive models developed for item generation serve as valid Q-matrices for CDM analyses in LSAs?

Since *ÉpStan*'s Grade 1 mathematics assessment is already based on developmental cognitive models (Sonnleitner et al., 2025), we test whether these models can be directly applied as Q-matrices without retrofitting. To evaluate how different conceptualizations influence CDM results, we compare two hierarchical models based on current theory:

- Single-attribute hierarchical model (SAHM): Each item is mapped to a single dominant attribute (counting, addition < 10, decomposition, or addition > 10), assuming a strict, sequential learning progression where more advanced skills can only be mastered once foundational skills are acquired. This conceptualization aligns with curriculum expectations and parallels Item Response Theory (IRT) in the sense that both frameworks assume a cumulative relation between ability and item difficulty (i.e. mastery of more complex skills presupposes mastery of simpler ones).
- Multiple-attribute hierarchical model (MAHM): This model allows interactions between attributes, acknowledging that students may apply multiple strategies when solving problems. Rather than assuming a rigid skill hierarchy, it captures a more flexible progression, where earlier-acquired skills may still be used alongside more advanced problem-solving techniques. This conceptualization aligns with previous findings on children's strategy use in arithmetic problem-solving (Siegler & Shrager, 1984).

What is the added value of CDMs in early-age LSAs?

Beyond feasibility, we examine whether CDMs can provide more fine-grained diagnostic insights at different levels of the analysis. Given the variability in early numeracy development – shaped by linguistic, cognitive, and sociocultural factors – we analyze how student background variables, such as the socioeconomic status, migration and language background, gender, and math-related attitudes, influence skill acquisition. By addressing these questions, the study aims to bridge the gap between system-level monitoring and classroom-level diagnostics, assessing whether CDMs can increase the instructional relevance of LSAs without compromising their core evaluation function.

Methods

Procedure

Data collection was conducted as part of the annual Luxembourgish school monitoring program, *ÉpStan* (Sonnleitner et al., 2018). The *ÉpStan* tests constitute Luxembourg's national large-scale assessment (LSA) program, assessing the full student cohort in grades 1, 3, 5, 7, and 9. Consequently, the *ÉpStan* represent the most comprehensive LSA possible in the Luxembourg context. In line with the National Research Council (Pellegrino, Chudowsky & Glaser, 2001) definition of LSAs, it serves a system-level monitoring function – providing data for evaluating curriculum implementation and equity across schools – rather than individual student certification. The *ÉpStan* achievement tests are based on the educational standards defined by the Ministry of Education, Children and Youth for all primary and secondary schools following the Luxembourgish curriculum (MENFP, 2011). The mathematic tests in primary schools retrospectively assess, at the beginning of a new learning cycle (i.e. grades 1, 3, 5), whether the educational goals of the previous learning cycle have been achieved by the students. The test items for Grade 1 were therefore developed according to two theoretical difficulty levels: Whereas theoretical level 1 („Niveau socle") items are quite easy and assess whether students achieved the minimum educational standards defined for Cycle 1 (preschool, e.g., counting a set of 5 objects), level 2 items are more difficult and assess the minimum educational standards defined for Cycle 2 (Grades 1 and 2; e.g., written additions in the number range 0 to 20).

Each first-grade student (the full cohort) completed two paper-pencil-based mathematics main test booklets and one of six randomly assigned pretest booklets, administered on separate days as part of the experimental design. Randomization occurred at the class level. Class teachers administered and guided students through the test booklets in a group setting, offering clarification when questions arose. Each booklet contained 20 to 45 items with varying response formats (open-ended, multiple choice, and matching) and was administered in class within a 30-minute session, guided in a standardized manner by the classroom teacher(s). The study received approval from the national Ministry of Education and adhered to national ethics and data protection regulations. Parents received individual feedback on their child's global mathematics performance, and teachers received both individual and class-level summaries. The results of the CDM analyses were not included in the feedbacks due to their exploratory nature.

Measures and design

Numeracy development

The primary goal of the ÉpStan mathematics test is to assess students' competencies in the curriculum-based domains of *Numbers & Operations* and *Space & Shape* (MENFP, 2011). This study focuses exclusively on subskills identified as key milestones in early numeracy development: counting, number and quantity decomposition, and symbolic addition below and above 10 (Krajewski & Schneider, 2009; Roesch & Moeller, 2015).

In the context of the *ÉpStan*, which monitor achievement after the mandatory two years of preschool and are administered in November (i.e., two months after the start of Grade 1), counting items range from 5 (expected) to 20 (ideal), while decomposition items and written addition, cover the number space from 0 to 20. Item examples are provided in Figure 1, and subskill descriptors are included in the mock feedback in Figure 2. Items were algorithmically generated using curriculum-aligned cognitive models. Each cognitive item model was directly derived from the national

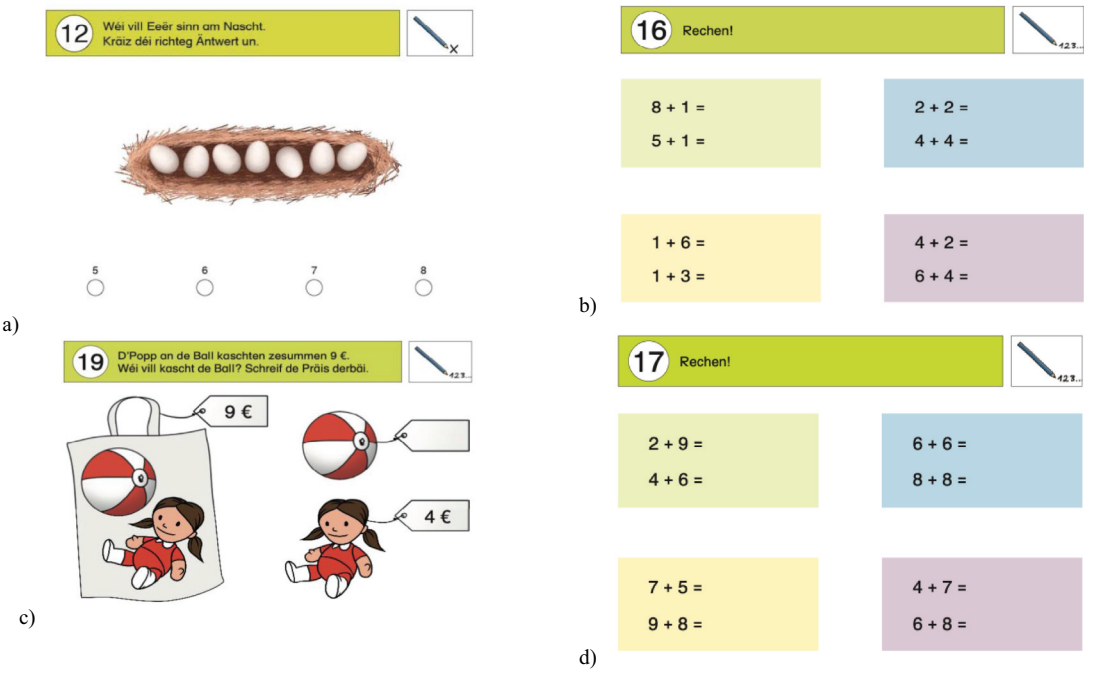


Figure 1. Item examples for each assessed numerical development stage: a) counting, b) addition < 10, c) decomposition, d) addition > 10.

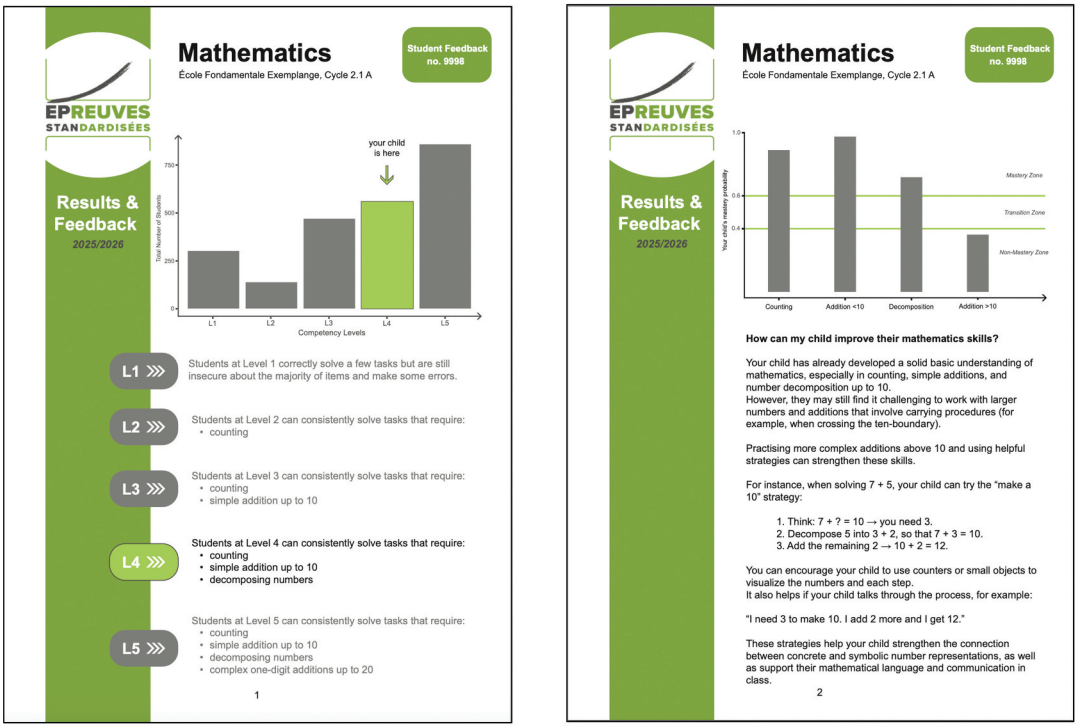


Figure 2. Example feedback for a student based on the final multiple-attribute hierarchical model (MAHM) giving information about the relative standing within the cohort and a more detailed individual evaluation of attribute mastery including strategies to improve.

curriculum standards for *Numbers & Operations* (MENFP, 2011) and defined the relevant cognitive attributes (e.g., counting, addition < 10, decomposition, addition > 10) as well as the permissible task variations. Using dedicated R-scripts, these parameters were systematically combined to generate multiple equivalent item instances for each attribute while preserving the intended cognitive demand. These cognitive item models and the corresponding Q-matrix were jointly developed by psychometricians, cognitive psychologists, and subject-matter experts (SMEs) in an iterative design process (see Sonnleitner et al., 2025). Each model explicitly defined the targeted competencies, cognitive processes, and attribute – item mappings at the time of construction, achieving consensus among all contributors. Consequently, Q-matrix validation was intrinsically embedded in the model development process and did not require a separate empirical or post-hoc validation step. The resulting item content characteristics were therefore clearly specified from the outset and directly informed the Q-matrix used in the CDM analyses.

For assessing early numeracy development, each student responded to 10 linkage items (8 included in the main test booklets, 2 included in the experimental pretest booklet) and 18 items that were specific to the administered pretest booklet. In total, 6 different pretest booklets were administered, which were randomly distributed at the class level. This corresponds to an incomplete design with linkage items, where missing values are completely at random (MCAR) and missing by design. Figure 1 shows example items for each subskill.

Students' demographic and sociocultural background characteristics

In addition to basic demographic variables such as age and gender, several key indicators known to influence mathematical development were collected through standardized student and parent questionnaires. Parents' responses on their current occupation were coded using the ISEI (International

Socio-Economic Index of Occupational Status; Ganzeboom et al., 1992). The highest ISEI (HISEI) of either parent was used as an indicator of the student's socioeconomic status (SES). Students were classified as being native if they themselves and at least one of their parents were born in Luxembourg. All other students were considered as having a migration background. Speaking the instructional languages Luxembourgish or German at home was considered advantageous. Students with such a language background were grouped as Germanophone and compared to their mainly French-speaking or Romance language-speaking peers (Portuguese-speaking children are considered a significant subgroup in Luxembourg's educational system with known disadvantages). To ensure adequate comparability across groups, only these largest and linguistically homogeneous home language groups (Germanophone, Francophone, and Romanophone) were retained. Students who reported speaking multiple languages at home (≈1,000) were excluded from these subgroup analyses, as the resulting small and heterogeneous groups would not allow for reliable estimation within the CDM framework. Accordingly, these subgroup comparisons should be regarded as exploratory illustrations rather than generalizable population estimates.

Students' motivational and affective factors

Questions referring to math-related anxiety, interest, and self-concept, as well as general school interest, were each asked by one dichotomous item for which students had to agree or disagree to a positively formulated statement.

Sample

In total, 2704 students (1313 girls, 10 did not report gender), nested in 209 classes of 82 schools, participated in the study. Testing took place in November, two months after the regular start of school. All students were enrolled in Grade 1 (Cycle 2.1) with an average age of 6.4 years (SD = 0.5 years). 804 students were mainly speaking Luxembourgish or German at home, 369 students French, and 482 students reported mainly speaking a Romance language. Note that students speaking more than one language at home ($n = 1049$) were omitted from language-specific analyses. Overall, 1329 students stated an immigration background (1st and 2nd generation). Classes were selected randomly, leading to a representative sample (45%) of the full cohort ($n = 5963$). Comparability to the full cohort was validated post-hoc concerning gender, age, language background, immigration and socioeconomic status.

Statistical analyses

Model specification and estimation

The Q-matrices used for CDM estimation were established a priori based on the cognitive models used during item generation. This ensured that attribute – item mappings were grounded in theory rather than derived empirically, minimizing the need for post-hoc modifications.

To evaluate and estimate the models, we first used the general G-DINA framework (De La Torre, 2011) to take advantage of its comprehensive set of item-level fit indices. The hierarchical models were then estimated using the Additive Cognitive Diagnostic Model (A-CDM; De La Torre, 2011), consistent with the compensatory nature of early numeracy skills. The A-CDM allows estimating compensatory (linear) relationships between attributes, where each attribute contributes independently to item success without higher-order interactions. This approach aligns with developmental models of numerical cognition (e.g. Krajewski & Schneider, 2009; Roesch & Moeller, 2015; von Aster, 2005) and simulation results indicating that interaction parameters add little beyond main effects (Kunina-Habenicht et al., 2012). Moreover, its parameter structure facilitates interpretable and instructionally actionable insights.

We first estimated a Single-Attribute Hierarchical Model (*SAHM*), assuming a theoretically grounded hierarchy among the four subskills: counting, addition < 10, decomposition, and addition > 10. Each

item was mapped to a single dominant attribute, supposing a strict, sequential learning progression where more advanced skills can only be mastered once foundational skills are acquired. This conceptualization aligns with curriculum expectations and parallels Item Response Theory (IRT) in the sense that both frameworks assume a cumulative relation between ability and item difficulty (i.e., mastery of more complex skills presupposes mastery of simpler ones). Given the theoretical validation of the Q-matrix, no modifications were made based on empirical estimates; however, items that showed poor model fit were removed, resulting in the refined *SAHM_r* model, which closely mirrored the item generation process. While many diagnostic modeling frameworks emphasize iterative refinement between theory and empirical validation (e.g., Tatsuoka et al., 2016), in the present study the Q-matrix was based on prevalidated cognitive item models developed with subject-matter experts. Additional empirical modification was intentionally limited to preserve theoretical coherence and avoid retrofitting.

Interpretation of item misfit drew on the theoretically specified and expert-validated cognitive item models used during test construction. Items that did not align empirically with their assigned attributes (most often those allowing multiple solution strategies) were therefore understood as reflecting expected developmental variability rather than flaws in item design. As children often use a variety of strategies when solving math problems – especially in early numeracy – a more flexible model was introduced. Prior research (e.g., Siegler, 1987, 1988; Siegler & Shrager, 1984) highlights that children draw on multiple strategies, such as counting, finger use, and retrieval, even for simple problems. Such flexibility, while developmentally adaptive, challenges single-attribute modeling. Particularly in less practiced tasks like decomposition, children may revert to more secure counting strategies (Carr & Jessup, 1997). Consequently, items inviting multiple solution routes are more prone to misfit under the *SAHM*.

To account for this, we estimated a Multiple-Attribute Hierarchical Model (*MAHM*), which allowed items to load on multiple attributes simultaneously, capturing the interdependence among early numeracy skills. Although the SAHM reflects a curriculum-based progression in which counting typically precedes decomposition and addition, empirical and developmental research shows that children use multiple strategies concurrently when solving number problems (e.g., counting all, counting min, decomposition, fact retrieval; Fuson, 1982; Siegler, 1987, 1988). The MAHM therefore represents a more flexible approach that accommodates this parallel strategy use by allowing partial mastery and compensatory interactions among skills. In this sense, the two models together capture both the sequential logic of curriculum-based expectations and the overlapping nature of cognitive development in early numeracy. The *MAHM* was subsequently refined by removing misfitting items, yielding the final *MAHM_r* model.

Model evaluation and fit assessment

To evaluate model appropriateness and ensure the validity of the estimated CDMs, we considered several established item fit statistics (e.g. De La Torre & Chiu, 2016; Kunina-Habenicht et al., 2009; Ma & de la Torre, 2020) and screened our item pool: a) RMSEA (root mean square error of approximation), with values < 0.07 was judged as moderate and therefore acceptable fit; b) MESA plots were examined to identify items that were not measuring the intended attributes using a pvaf (proportion of variance accounted for each q-vector specified in the q-matrix) lower than 0.95; c) G-DINA discrimination index (GDI), which measures the variance of item success probabilities based on the reduced attribute profile, with items below .01 flagged for review; and d) P(1) – P(0) (discrimination power) between .20 and .75 to identify weakly discriminating items. Finally, we examined items showing e) an unusually high probability of a correct response among students who had not mastered any of the required attributes ($g > .70$) or f) a high probability of incorrect response among those who had mastered at least one required attribute ($s > .50$), and were thus considered misfitting and removed. These thresholds follow established practice in CDM applications (e.g. De La Torre & Chiu, 2016) and are intended to flag items whose response patterns deviate substantially from their theoretical cognitive specifications.

This statistical screening complemented the prior content validation carried out during cognitive model development with SMEs, ensuring alignment between theoretical assumptions and empirical item behavior. Note that while a unidimensional baseline model could theoretically serve as a reference point, our focus was to evaluate whether different hierarchical cognitive structures (*SAHM* vs. *MAHM*) meaningfully captured expected developmental patterns rather than to test overall dimensionality.

Following the selection of the best-fitting models, CDM results were examined at three levels: (1) cohort-level, describing overall mastery distributions and model fit; (2) group-level, analyzing differences in the distribution of latent mastery profiles across subgroups defined by gender, SES, language background, and motivational-affective indicators (Group-level comparisons were based on observed and expected frequencies within each latent class, with statistical significance indicated in Figures 4 and 5); and (3) individual-level, illustrated through a mock student feedback example derived from posterior mastery probabilities to demonstrate how CDM-based profiles could inform personalized reporting. The cohort- and group-level analyses are reported in Numeracy development at the beginning of grade 1 on the subgroup level. All items included in the analysis underwent routine differential item functioning (DIF) screening as part of the *ÉpStan*'s annual technical quality assurance (Fischbach et al., 2014), ensuring that subgroup comparisons were based on psychometrically verified items.

All analyses were conducted in R (R Core Team, 2023). Model estimation and primary diagnostics were conducted using the GDINA package (Ma & de la Torre, 2020). To obtain additional fit statistics not directly available in GDINA, the estimated parameters were fixed and evaluated using functions from the CDM package (George et al., 2016), ensuring that identical model specifications were assessed across packages.

Results

Model fit and measurement accuracy

Single-attribute hierarchical model (SAHM)

The initial *SAHM* comprised of 95 items of which 28 items were dropped based on item fit criteria mentioned above. Content analysis showed that the majority of lost items belonged to item models that potentially allowed for different solving strategies (e.g. either addition or decomposition items) indicating that the Single-Attribute model based on primary skill measured may be too strict or narrow in its conceptualization. The remaining 67 items nevertheless provided a good coverage of the tested attributes (see Table 1) and the resulting reduced model *SAHM_r* showed good fit and acceptable classification accuracy of .85. The distribution of items across attributes and their classification accuracies are summarized in Table 1, which provides an overview of the test composition. Table 2 shows absolute and relative fit statistics.

Multiple attribute hierarchical model (MAHM)

Taking the high loss of items in the *SAHM* into consideration, we further investigated a multiple attribute hierarchical model that accounts for items measuring several attributes which are organized in a hierarchical way. The underlying Q-matrix resulted out of 6 iterations that were based on

Table 1. Number of items per competency and attribute level accuracy in the final estimated models.

	Counting	Addition < 10	Decomposition	Addition > 10	Total
SAHM_r					
Items per attribute	9	25	15	18	67
Attribute level accuracy	.96	.96	.94	.95	
MAHM_r					
Items per attribute	10	28	31	20	89
Attribute level accuracy	.96	.97	.95	.96	

Table 2. Absolute and relative model fit statistics for (revised) single-attribute hierarchical model (*SAHM_r*) and (revised) multiple-attribute hierarchical model (*MAHM_r*).

	SAHM	SAHM_r	MAHM	MAHM_r
RMSEA2	0.054	0.038	0.035	0.052
SRMSR	0.050	0.050	0.041	0.050
MADQ3	0.052	0.051	0.049	0.051
Loglikelihood	−33511.66	−26522.14	−33149.29	−32297.94
AIC	67,023.33	53,044.30	66,298.59	64,595.89
BIC	67,023.33	53,044.30	66,298.59	64,595.89
Accuracy	0.87	0.85	0.87	0.86

Table 3. Pattern level accuracy for (revised) single-attribute hierarchical model (*SAHM_r*) and (revised) multiple-attribute hierarchical model (*MAHM_r*).

Pattern	*SAHM_r*	*MAHM_r*
0000	.74	.79
1000	.78	.68
1100	.86	.86
1110	.86	.86
1111	.95	.94

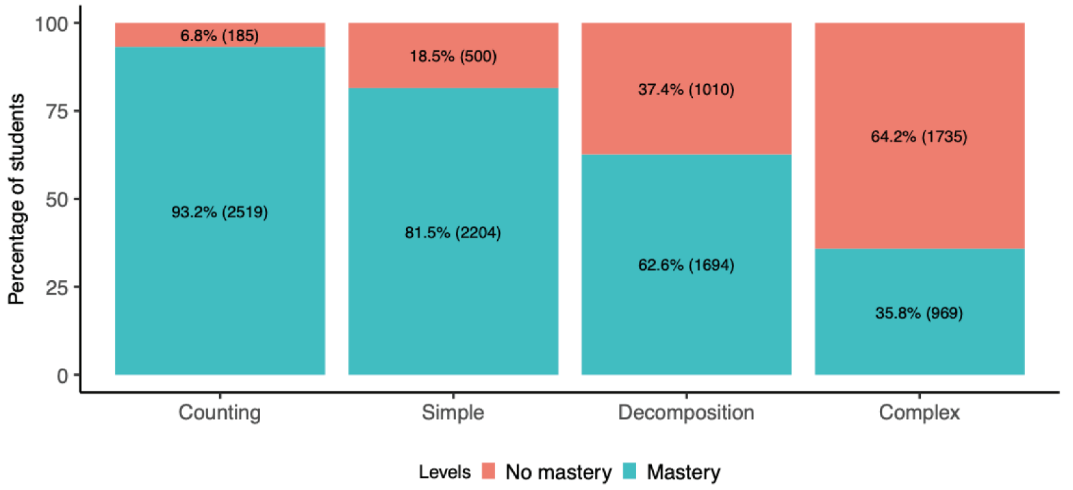
modification indices which were iteratively discussed and approved by SMEs. Through this procedure, only six out of the initial 95 items were lost due to violating item fit indices. The resulting model *MAHM_r* showed comparable fit and measurement accuracy (.86) to the *SAHM_r*. Note that the biggest difference to the *SAHM_r* lies in better accommodating items measuring decomposition – items which could probably be solved using different strategies (e.g. counting). Attribute-level accuracy ranged between .95 and .97. Pattern-level accuracy was comparable between *SAHM_r* and *MAHM_r* and ranged for the latter from .68 for the latent class only mastering the first attribute (1000) to .94 for the latent class mastering all attributes (1111), see Table 3. Pattern-level accuracy was clearly related to a lower number of items measuring counting compared to the other attributes.

In sum, both conceptualizations showed comparable global fit statistics within an acceptable to good range and demonstrated satisfactory measurement accuracy. Because the *SAHM* and *MAHM* differed slightly in their retained item sets, direct statistical comparison of global fit indices was not feasible. Therefore, equivalence in fit was judged based on the pattern and magnitude of multiple indices (e.g., RMSEA, pvaf, GDI), supported by the theoretical interpretability of both models. These findings suggest that model-based item development offers a promising foundation for applying CDMs, even in originally non-diagnostic LSA contexts. However, modeling each item with a single attribute proved challenging, particularly for tasks that allow multiple solution strategies – a result anticipated both from statistical considerations and theoretical models of number development. Consequently, we focussed on the refined Multiple-Attribute Hierarchical Model (*MAHM_r*) for subsequent analyses at the cohort, subgroup, and individual levels. Substantial deviations from the refined Single-Attribute Hierarchical Model (*SAHM_r*) will be explicitly noted.

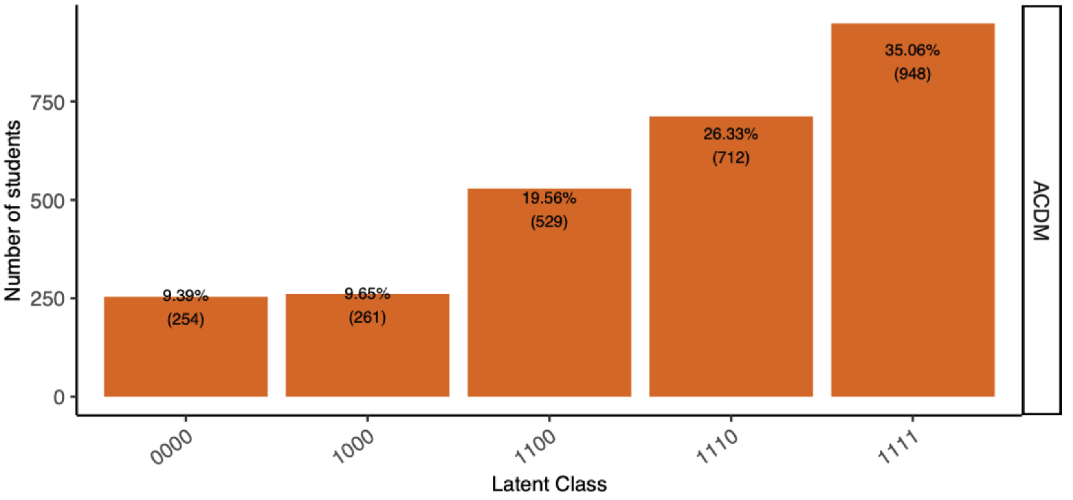
Numeracy development at the beginning of grade 1 on the cohort level

Cohort-level results aligned with expectations from developmental models (see Figure 3(a)): as attribute complexity increased, the proportion of students mastering each skill decreased. Whereas *counting objects* was mastered by 93.2% of the students, *addition < 10* by 81.5%, and *decomposition problems* by 62.6%, only 35.8% of students reliably succeeded in solving *addition problems > 10*.

This increasing difficulty is reflected when looking at the student distribution by latent classes (Figure 3(b)). Note that the applied hierarchical model *MAHM_r* only allowed for theoretically meaningful classes: a higher skill could only be mastered if the attributes below are already mastered.



a) Percentage of students mastering number development skills according to the *MAHM_r* model



b) Percentage of students within distinct latent classes according to the *MAHM_r* model

Figure 3. Numeracy development at the beginning of grade 1 on the cohort level.

Roughly a third of the students (35.06%) mastered all tested skills and have therefore successfully passed all steps in number development as tested within the *ÉpStan*. The rest of the students (55.54%) can be differentially classified to either mastering one, two or three of the four tested skills. Note that the lower pattern level accuracy for class 1000 (.68) might have led to a slightly imprecise categorization of students at the lower end of the proficiency distribution. More dedicated items might help with this.

These findings confirm that it is feasible to differentiate students' numerical development within Grade 1 in an LSA setting. Notably, although mastering complex additions (>10) is not formally expected at the time of testing, a substantial proportion of students (35.8%) had already acquired it. This highlights the variability in early mathematical learning and suggests that solving complex additions may serve as an early indicator of advanced numerical proficiency.

However, while the observed mastery patterns broadly align with theoretical expectations of early numerical development (see above), they may also reflect differences in item representation or sampling across attributes (e.g., variation in task format or visual embedding). Therefore, interpretations should be made cautiously and in light of the specific item characteristics included in the current design.

Numeracy development at the beginning of grade 1 on the subgroup level

To better understand students' numeracy development and assess the added value of CDMs, latent classes were analyzed in relation to students' background characteristics. Figures 4 and 5 show the percentage over- and underrepresentation of different subgroups across latent classes, offering deeper insight into how background variables relate to numeracy development.

Gender

Girls were slightly overrepresented in all developmental classes except the highest latent class (1111). While 41.3% of boys demonstrated mastery of all tested attributes, only 28.5% of girls reached this level. These results offer a more detailed picture of the gender differences previously reported in the *ÉpStan* (e.g., Sonnleitner et al., 2018), indicating that such disparities emerge early in the development of core numerical skills. They may also reflect differing learning trajectories or arithmetic strategies: prior studies have shown that girls are more likely to rely on overt strategies like counting, while boys tend to use mental strategies such as decomposition and retrieval (Carr & Jessup, 1997).

Socioeconomic status

When analyzing students based on socioeconomic status (SES), clear disparities in numerical development were observed across all latent classes. Students in the lowest HISEI quartile (Q1) were overrepresented in lower developmental stages, particularly in early numerical skills (e.g., counting and addition < 10). Nearly half (46.92%) of high-SES students (Q4) mastered all tested developmental steps, whereas this was true for only 24.32% of low-SES students (Q1).

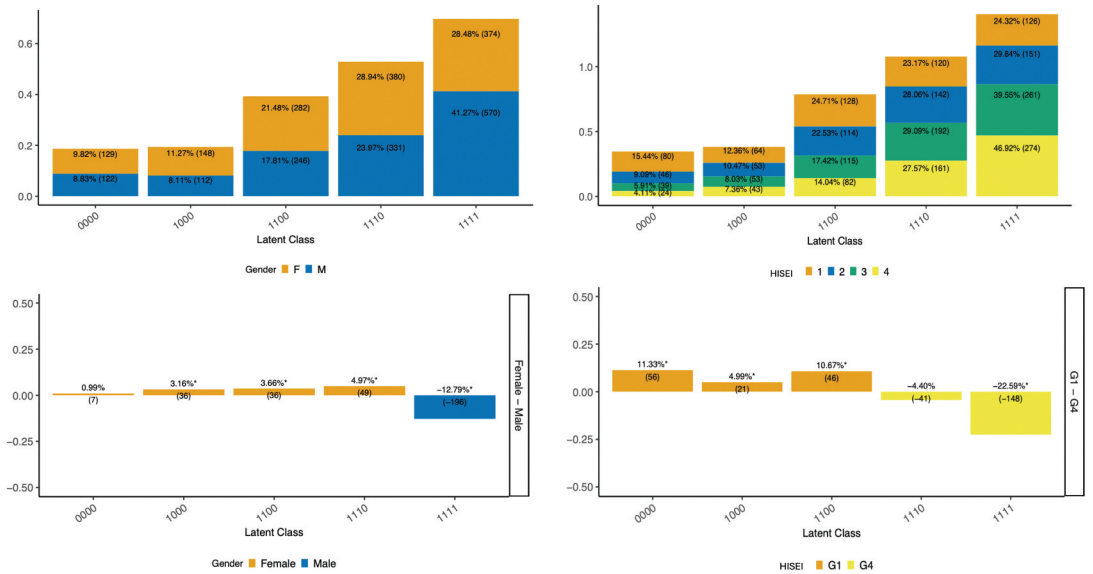


Figure 4. Proportion of students per latent class (top) and difference in proportion of students (bottom) per subgroup (gender left and SES right) based on the *Mahm_r* model. *indicates a statistically significant difference.

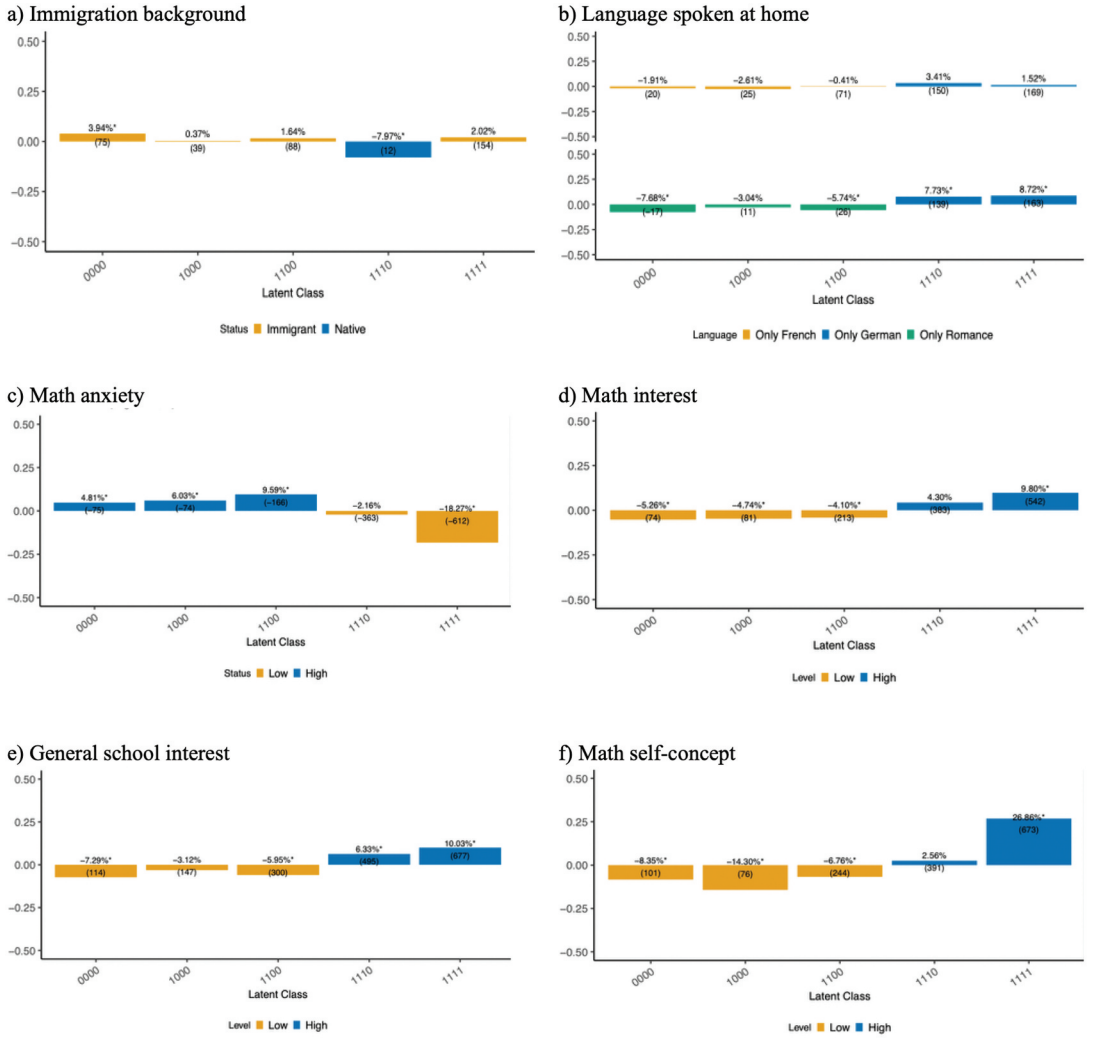


Figure 5. Difference in proportion of students (bottom) per subgroup (gender left and SES right) per latent class based on the *Mahm_r* model. *indicates a statistically significant difference.

Although SES-related differences were already evident in the earliest developmental stages, they became even more pronounced after mastering decomposition. The gap between Q1 and Q2, as well as Q3 and Q4, widened further at this stage, suggesting that while lower-SES students initially lag behind, the discrepancy in skill acquisition accelerates as mathematical complexity increases. Importantly, compared to gender differences, SES-related disparities appear at even earlier stages of numerical development. While 15.4% of Q1 students struggled with basic counting, only 4.1% of Q4 students exhibited similar difficulties. This suggests that SES-related performance gaps in mathematics emerge early in schooling and continue to expand, underscoring the importance of early intervention strategies to support students from disadvantaged backgrounds (cf. Jordan et al., 2010; Nelson & Powell, 2018).

Immigration background

Performance differences based on migration background were relatively small compared to gender and SES-related disparities. This might be partly explained by the formal/high preschool attendance

which is compulsory for two years at the age of 4 (plus one facultative for 3-to-4-year olds) where pre-math teaching (numeracy) are an essential part of the curriculum preparing the kids for primary school (cf. ECEC report)

A slightly higher proportion of non-native students (+3.9%) struggled with all measured attributes, while 7.9% fewer reached mastery of decomposition. These findings suggest that while migration background is associated with some variation in early numerical development, its impact appears less pronounced than other background characteristics, such as socioeconomic status.

Language spoken at home

No significant performance differences were found between students who primarily spoke the instructional languages (Luxembourgish or German) and those who spoke French. In contrast, students from Romanophone-speaking households (mostly Portuguese) were overrepresented in the lowest latent classes (0000: +7.7%, 1100: +5.7%) and underrepresented in the higher ones (1110: −7.7%, 1111: −8.7%). Language skills are essential for learning the number word sequence ("one, two, three, four . . . "; Fuson, 1988) and for developing counting and arithmetic skills (Donlan et al., 2007). Furthermore, children's home language background plays a crucial role in mathematics development. They tend to calculate more effectively in the language in which they first learned numbers – usually their first language (Geary et al., 1993). Arithmetic skills represent learned content that is not easily transferable from one language to another without cognitive costs (Saalbach et al., 2013; Spelke & Tsivkin, 2001). In Luxembourg, where more than 65% of students learn mathematics in a language different from their home language, recent studies have shown that home language predicts mathematics performance, with students not speaking the instructional language at home performing lower than their peers who do (Greisen et al., 2021; Hornung et al., 2021). However, these relationships are correlational and partly reflect socioeconomic disparities between language groups.

Motivational and affective factors in numeracy development

The relationship between mathematical competency and affective factors, such as math anxiety, interest, and self-concept, is well-documented in the literature. The present findings further highlight how these factors interact with early numeracy development.

A clear inverse relationship emerged between math anxiety and numeracy development. Students with high anxiety were increasingly overrepresented in lower latent classes, peaking among those who had mastered only two developmental stages (1100). After mastering decomposition (1110), anxiety levels stabilized, but the highest class (1111) showed a marked overrepresentation of low-anxiety students (+18.3%). However, 22.1% of students who reported high math anxiety mastered all attributes, suggesting that anxiety does not vanish even at advanced levels. Overall, math anxiety decreased with proficiency but followed a non-linear trajectory, peaking at intermediate stages before dropping – a pattern that may reflect students' struggle with increasing task complexity before confidence solidifies with full mastery.

A small but statistically significant trend was found for math interest, with students reporting low interest consistently overrepresented in lower latent classes up to 1100. From decomposition mastery onward (1110), students with high math interest became increasingly overrepresented, with the strongest effect in the highest class (1111: +9.8%).

A similar, though weaker, pattern was found for general school interest, suggesting that domain-specific motivation plays a more pronounced role in mathematical skill development than general academic engagement.

Mathematical self-concept showed the most striking pattern. Up to decomposition mastery (1110), students were significantly more likely to report low self-concept (0000: −8.4%, 1000: −14.3%, 1100: −6.8%). No significant differences were found at 1110, but in the highest latent class (1111), there was a strong overrepresentation of students with high self-concept (+26.9%). The proportion of students with high self-concept increased from 14.29% in lower classes to 41.14% in the highest class. This pattern likely reflects that students who have mastered decomposition are more aware of their

mathematical strengths, whereas inflated self-perceptions in lower classes may stem from competencies not yet acquired.

These results suggest that mastering decomposition marks a pivotal transition – not only in mathematical competence but also in students' self-perception. The sharp rise in self-concept among students who mastered all skills underscores the close link between competence and confidence. Math anxiety decreases only after full mastery, while interest and self-concept increase notably. Supporting students through this stage appears crucial, as it strongly shapes both their cognitive development and affective engagement in mathematics.

Discussion

This study investigated the feasibility and added value of Cognitive Diagnostic Models (CDMs) in a large-scale educational assessment (LSA) setting, using the Luxembourg's national school monitoring program *ÉpStan* as a case study. Our approach leveraged cognitive models originally developed for item generation, examining their direct applicability as Q-matrices for CDM analyses without requiring post hoc modifications.

Findings suggest that cognitive models effectively structured Q-matrices, allowing for meaningful skill assessment without retrofitting. Both the Single-Attribute Hierarchical Model (*SAHM*) and the Multiple-Attribute Hierarchical Model (*MAHM*) successfully replicated expected developmental progressions in early numeracy, with decomposition proficiency emerging as a critical threshold with regard to students' background variables. This aligns with prior research emphasizing the essential role of decomposition in first-grade arithmetic for future mathematics learning (Geary et al., 2013; Laski et al., 2014). The higher item attrition in the *SAHM* likely reflects genuine overlap among early arithmetic skills rather than measurement flaws. Decomposition items are inherently open to alternative strategies (e.g., counting or retrieval), making strict one-attribute mappings unrealistic. This underscores the need for modeling frameworks that accommodate strategy variability (e.g., *MAHM*) and for curricular designs that encourage explicit strategy instruction to help children transition from concrete counting toward decomposition-based reasoning (cf. Carr & Jessup, 1997). In this regard, the relatively high item attrition under the *SAHM* should not be interpreted as a weakness of model-based item development but rather as a reflection of the strong theoretical constraints imposed by single-attribute modeling. When alternative strategies are possible, such as in decomposition tasks, items may deviate from a strictly hierarchical structure even though they remain cognitively valid. The higher retention under the *MAHM* supports that cognitive modeling provided a transparent framework for understanding and interpreting such misfit.

The developmental patterns identified in this study, particularly the emergence of decomposition as a pivotal threshold, align with broader evidence that mathematical development follows hierarchically ordered progressions (e.g., Geary et al., 2013; LeFevre et al., 2010; Lyons et al., 2014; M. G. Von Aster & Shalev, 2007). Early number development is grounded in a preverbal sense of quantity (Butterworth, 2005; Dehaene, 1997), which gradually links to symbolic number systems and arithmetic through counting and number comparison skills (Aunola et al., 2004; Geary, 2013; Hornung et al., 2014; Jordan et al., 2010). While these processes are assumed to follow similar trajectories across cultures (Vasilyeva et al., 2023), their pace and manifestation differ depending on curricular emphases, language transparency, and early education systems (Dowker et al., 2008; Mark & Dowker, 2015). Consequently, our results likely reflect universal cognitive mechanisms embedded in the specific multilingual and curriculum-aligned context of Luxembourg. Caution is warranted when generalizing to educational systems with differing instructional sequencing or linguistic number structures.

Subgroup analyses provided valuable insights into individual differences in numeracy development. SES-related disparities emerged early, confirming findings that mathematical inequalities manifest before formal schooling begins (Jordan et al., 2010; Nelson & Powell, 2018). The small differences observed by migration background are consistent with the nature of the assessment in Luxembourg's early education context. The *ÉpStan* tests in Grade 1 use very short instructions and are

administered with teacher guidance (e.g. teachers may translate the instructions). The items focus on basic numerical skills, such as counting or comparing numbers, which children can solve in their familiar (first) language. Moreover, most children (regardless of migration background) have completed at least two years of early childhood education in Luxembourgish before school entry, which supports the comprehension of task instructions and reduces language-related differences (cf. Hornung et al., 2023). As mathematical content becomes more verbally and conceptually demanding in later grades, migration-related gaps may become more pronounced (for the Luxembourgish context, see e.g. Sonnleitner et al., 2018). Analyzing affective factors, we observed a nonlinear relationship between math anxiety and skill level, with anxiety peaking among students at intermediate developmental stages – a pattern consistent with theories on competence-related stress in learning (Dowker et al., 2016).

Beyond group-level comparisons, CDM-based classifications also enable fine-grained feedback at the individual student level. Figure 2 illustrates an example of such feedback based on the final *MAHM*, showing a student's relative standing within the cohort and a detailed evaluation of their mastery across cognitive attributes. However, in the current study, such feedback was used illustratively rather than operationally, as the focus remained on evaluating the feasibility of the CDM approach within an LSA framework. Nevertheless, it demonstrates how CDM results could be translated into meaningful feedback for teachers and students, informing targeted instructional strategies and individualized support. Such individual profiles should be interpreted as probabilistic summaries derived from population-calibrated models at a given assessment occasion, rather than as direct representations of intraindividual learning processes or change. In line with methodological discussions on individual inference from interindividual models (e.g., Molenaar, 2004), CDM-based feedback is best understood as structured decision support for instruction grounded in cognitively defined skill profiles, rather than as a model of intraindividual learning dynamics or moment-to-moment cognitive processes.

Together, these findings highlight that CDMs grounded in cognitive models which were used for item generation offer significant potential to enhance LSAs. By embedding structured developmental frameworks into assessment design, LSAs could serve not only as system-level monitoring tools but also as sources of meaningful, instructionally relevant feedback. However, realizing this potential depends not only on technical advances in assessment design but also on teachers' pedagogical content knowledge about how mathematical skills develop, which shapes how they interpret and act upon feedback (Ban et al., 2024). At the same time, it is important to acknowledge that CDMs and LSAs operate at different levels of granularity and pursue distinct purposes. Whereas LSAs primarily provide population-level indicators of achievement and equity, CDMs are designed for fine-grained diagnosis at the level of individual learners. In large-scale contexts, item sampling and testing-time constraints limit the reliable estimation of detailed mastery profiles for every student. Nevertheless, integrating cognitive models into LSAs can enhance their interpretability and instructional relevance by identifying typical learning progressions and common developmental bottlenecks observable at scale. Thus, rather than replacing traditional LSA functions, CDMs can complement them – offering a conceptual bridge between system-level monitoring and classroom-level insight.

Despite its contributions, the study has several limitations. First, while our findings suggest that cognitive models can directly structure Q-matrices, cross-validation in independent samples would be needed to confirm generalizability. Future studies should examine the stability of these models across cohorts, cultural contexts, and later mathematical competencies to assess their viability across different domains. Second, the applied setting of the study led to a design including many missings. Applying the same Q-matrix structure to full data might allow for better estimating the results' robustness. Another limitation concerns the assessment of affective – motivational constructs. Math anxiety, interest, and self-concept were each measured using single dichotomous items. While this approach restricts reliability and construct validity, it was necessary given the young age of the participants and the large-scale assessment context, where testing time and response formats must remain age-appropriate. Previous research indicates that ultra-short and even single-item scales can provide

psychometrically sound and practically valid estimates of motivational constructs when longer instruments are not feasible (Gogol et al., 2014). Accordingly, these variables should be interpreted as exploratory indicators that capture broad tendencies rather than precise trait estimates.

A final methodological consideration concerns the planned missingness design of the pretest booklets. While this incomplete design was implemented to efficiently sample a wide range of items, it may, in principle, influence classification accuracy. However, additional analyses confirmed that booklet-level differences in item difficulty were negligible and that only small variations in student ability distributions occurred across booklets. A full reanalysis of CDM results per booklet was not conducted, as dividing the sample into six smaller groups would have resulted in insufficient case numbers per latent class for stable estimation. In the full cohort, class sizes and model fit statistics indicated robust classification, suggesting that the impact of the planned missingness design on the results is minimal.

Additionally, while our study successfully differentiated skill levels at the cohort and subgroup levels, the practical implications for individual student feedback warrants further investigation. Validity evidence concerning individual feedbacks resulting from this study and their reception among educators should be investigated to fully ensure the transfer and added value of CDM insights into instructional practice. In this regard, further professional development programs focused on teachers' assessment literacy to ensure leveraging CDM insights for instructional decisions would be recommended.

Our findings suggest that using cognitive models in LSAs bridges the gap between system monitoring and instructional feedback. By integrating developmental frameworks directly into the test design, CDMs offer a scalable approach for enhancing the instructional relevance of LSAs without compromising their large-scale monitoring functions. Related work applying similar cognitive models to Grade 3 data from Luxembourg's national mathematics LSA and focusing on skill hierarchies (Effatpanah et al., 2026) provides complementary evidence that the proposed framework extends beyond early numeracy and supports its broader applicability. Together, these findings demonstrate a pathway toward more meaningful assessment practices, where LSAs can serve both policymakers and teachers alike.

Acknowledgments

We thank Cécile Braun for her careful proofreading of the manuscript.

Disclosure statement

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

Funding

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

References

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *96*, 699–713. https://doi.org/10.1037/0022-0663.96.4.699

Ban, J., Msall, C., Douglas, A. A., Rittle-Johnson, B., & Laski, E. V. (2024). Knowing what they know: Preschool teachers' knowledge of math skills and its relation to instruction. *Journal of Experimental Child Psychology*, *246*, 105996. https://doi.org/10.1016/j.jecp.2024.105996

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement Issues & Practice*, *18*, 5–12. https://doi.org/10.1111/j.1745-3992.1999.tb00266.x

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*, 371–399. https://doi.org/10.1146/annurev.psych.53.100901.135233

Braeuning, D., Hornung, C., Hoffmann, D., Lambert, K., Ugen, S., Fischbach, A., Schiltz, C., Hübner, N., Nagengast, B., & Moeller, K. (2021). Long-term relevance and interrelation of symbolic and non-symbolic abilities in mathematical-numerical development: Evidence from large-scale assessment data. *Cognitive Development*, *58*. https://doi.org/10.1016/j.cogdev.2021.101008

Butterworth, B. (2005). The development of arithmetic abilities. *Journal of Child Psychology and Psychiatry*, *46*(1), 3–18. https://doi.org/10.1111/j.1469-7610.2004.00374.x

Carr, M., & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy use: Social and metacognitive influences. *Journal of Educational Psychology*, *89*(2), 318. https://doi.org/10.1037/0022-0663.89.2.318

Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *EURASIA Journal of Mathematics, Science and Technology Education*, *11*, 1563–1577. https://doi.org/10.12973/eurasia.2015.1421a

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*, 415–427. https://doi.org/10.1016/j.econedurev.2008.09.003

Clark, A. K., & Karvonen, M. (2020). Constructing and evaluating a validation argument for a next-generation alternate assessment. *Educational Assessment*, *25*(1), 47–64. https://doi.org/10.1080/10627197.2019.1702463

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*, e2281. https://doi.org/10.1002/icd.2281

Davis-Kean, P. E., Tighe, L. A., & Waters, N. E. (2021). The role of parent educational attainment in parenting and children's development. *The Current Directions in Psychological Science*, *30*(2), 186–192. https://doi.org/10.1177/0963721421993116

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press. https://doi.org/10.1017/s0790966700006248

De Lange, J. (2007). Large-scale assessment and mathematics education. *Second Handbook of Research on Mathematics Teaching and Learning*, *2*, 1111–1144.

De La Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183. https://doi.org/10.1177/0146621608320523

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

De La Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273. https://doi.org/10.1007/s11336-015-9467-8

DeLuca, C., Schneider, C., Coombs, A., Pozas, M., & Rasooli, A. (2020). A cross-cultural comparison of German and Canadian student teachers' assessment competence. *Assessment in Education Principles, Policy & Practice*, *27*(1), 26–45. https://doi.org/10.1080/0969594x.2019.1703171

Donlan, C., Cowan, R., Newton, E. J., & Lloyd, D. (2007). The role of language in mathematical development: Evidence from children with specific language impairments. *Cognition*, *103*, 23–33. https://doi.org/10.1016/j.cognition.2006.02.007

Dowker, A., Bala, S., & Lloyd, D. (2008). Linguistic influences on mathematical development: How important is the transparency of the counting system? *Philosophical Psychology*, *21*(4), 523–538. https://doi.org/10.1080/09515080802285511

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, *7*, 508. https://doi.org/10.3389/fpsyg.2016.00508

Effatpanah, F., Kunina-Habenicht, O., Bernard, S., Hornung, C., & Sonnleitner, P. (2026). Optimizing large-scale mathematical assessments: Leveraging hierarchical attribute structures and diagnostic classification models for enhanced student diagnostics. *Educational Measurement Issues & Practice*.

Eisenwort, B., Aslan, H., Yesilyurt, S. N., Till, B., & Klier, C. M. (2018). Sprachentwicklung bei Kindern mit Migrationshintergrund und elterliches Vorlesen [Language development in children with migration background and parental reading to children]. *Zeitschrift Fur Kinder- Und Jugendpsychiatrie Und Psychotherapie*, *46*(2), 99–106. https://doi.org/10.1024/1422-4917/a000500

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

Fischbach, A., Ugen, S., & Martin, R. (2014). Épstan technical report. University of Luxembourg, LUCET. https://orbilu.uni.lu/bitstream/10993/15802/1/%c3%89pStan%20Technical%20Report.pdf

Fuson, K. C. (1982). An analysis of the counting-on solution procedure in addition. In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 67–81). Erlbaum. https://doi.org/10.4324/9781003046585

Fuson, K. C. (1988). *Children's counting and concept of number*. Springer. https://doi.org/10.1007/978-1-4612-3754-9

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

Geary, D. C. (2013). Early foundations for mathematics learning and their relations to learning disabilities. *The Current Directions in Psychological Science*, *22*(1), 23–27. https://doi.org/10.1177/0963721412469398

Geary, D. C., Cormier, P., Goggin, J. P., Estrada, P., & Lunn, M. C. E. (1993). Mental arithmetic: A componential analysis of speed-of-processing across monolingual, weak bilingual, and strong bilingual adults. *International Journal of Psychology*, *28*(2), 185e201. https://doi.org/10.1080/00207599308247184

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLOS ONE*, *8*, e54651. https://doi.org/10.1371/journal.pone.0054651

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*, 1–24. https://doi.org/10.18637/jss.v074.i02

Gersten, R., & Chard, D. (1999). Rethinking arithmetic instructions for students with mathematical disabilities. *Journal of Special Education*, *33*(1), 18–28. https://doi.org/10.1177/002246699903300102

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*, 318–341. https://doi.org/10.1080/15305058.2010.509554

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 263–268. https://doi.org/10.1080/15366360802497762

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My questionnaire is too long!" the assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, *39*, 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

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

Haghayeghi, M., Moghadamzadeh, A., Ravand, H., Javadipour, M., & Kareshki, H. (2025). Development of a cognitive assessment checklist for first-grade mathematics: Utilizing hierarchical cognitive diagnostic modeling in elementary Education. *Journal of Psychoeducational Assessment*, *43*(1), 88–107. https://doi.org/10.1177/07342829241290277

Holliday, W. G., & Holliday, B. W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Educational Forum*, *67*(3), 250–258. https://doi.org/10.1080/00131720309335038

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Hornung, C., Kaufmann, L. M., Ottenbacher, M., Weth, C., Wollschläger, R., Ugen, S., & Fischbach, A. (2023). *Early childhood education and care in Luxembourg. Attendance and associations with early learning performance*. Luxembourg Center of Educational Testing (LUCET). https://doi.org/10.1080/10409289.2025.2577955

Hornung, C., Schiltz, C., Brunner, M., & Martin, R. (2014). Predicting first-grade mathematics achievement: The contributions of domain-general cognitive abilities, nonverbal number sense, and early number competence. *Frontiers in Psychology*, *5*, 1–17. https://doi.org/10.3389/fpsyg.2014.00272

Hornung, C., Wollschläger, R., Keller, U., Esch, P., Muller, C., & Fischbach, A. (2021). Neue längsschnittliche Befunde aus dem nationalen Bildungsmonitoring ÉpStan in der 1. und 3. In LUCET. & SCRIPT. (Eds.), *Klasse: Negativer Trend in derKompetenzentwicklung und kein Erfolg bei Klassenwiederholungen* (pp. 44–55). LUCET & SCRIPT.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge University Press. https://doi.org/10.1017/cbo9780511611186.002

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, *20*, 82–88. https://doi.org/10.1016/j.lindif.2009.07.004

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*(3), 850–867. https://doi.org/10.1037/a0014939

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning & Instruction*, *19*, 513–526. https://doi.org/10.1016/j.learninstruc.2008.10.002

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*, 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2017). Incremental validity of multidimensional proficiency scores from diagnostic classification models: An illustration for elementary school mathematics. *International Journal of Testing*, *17*, 277–301. https://doi.org/10.1080/15305058.2017.1291517

Laski, V. L., Ermakova, A., & Vasilyeva, M. (2014). Early use of decomposition for addition and its relation to base-10 knowledge. *Journal of Applied Developmental Psychology*, *35*(5), 444–454. https://doi.org/10.1016/j.appdev.2014.07.002

LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, *81*(6), 1753–1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press. https://doi.org/10.1017/cbo9780511611186

Li, L., Zhou, X., Gao, X., & Tu, D. (2020). The development and influencing factors of kindergarteners' mathematics problem solving based on cognitive diagnosis assessment. *ZDM*, *52*, 677–690. https://doi.org/10.1007/s11858-020-01153-x

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from irt-based assessment forms. *Educational and Psychological Measurement*, *78*, 357–383. https://doi.org/10.1177/0013164416685599

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1-6. *Developmental Science*, *17*(5), 714–726. https://doi.org/10.1111/desc.12152

Ma, W., & de la Torre, J. (2020). GDINA: An R Package for Cognitive Diagnosis Modeling. Journal of Statistical Software, 93(14), 1–26. https://doi.org/10.18637/jss.v093.i14

Mark, W., & Dowker, A. (2015). Linguistic influence on mathematical development is specific rather than pervasive: Revisiting the Chinese number advantage in Chinese and English children. *Frontiers in Psychology*, *6*, 203. https://doi.org/10.3389/fpsyg.2015.00203

Mazzocco, M. M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research and Practice*, *20*(3), 142–155. https://doi.org/10.1111/j.1540-5826.2005.00129.x

MENFP. (2011). *Kompetenzraster und entwicklungsstufen. Grundschule, Zyklen 1 bis 4*.

MENJE & SCRIPT. (2024). *Education system in Luxembourg. Key figures. School year 2023/2024*. SCRIPT.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Muñez, D., Bull, R., & Lee, K. (2021). Socioeconomic status, home mathematics environment and math achievement in kindergarten: A mediation analysis. *Developmental Science*, *24*(6), e13135. https://doi.org/10.1111/desc.13135

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/

National Council of Teachers of Mathematics. (2017). Catalyzing change in high school mathematics. https://www.nctm.org/uploadedFiles/Standards_and_Positions/CatalyzingChangePublichReview.pdf

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of Learning Disabilities*, *51*(6), 523–539. https://doi.org/10.1177/0022219417714773

OECD. (2024). *Education at a glance, 2024: OECD indicators*. https://doi.org/10.1787/c00cad36-en

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.

Platas, L., Ketterlin-Gellar, L., Brombacher, A., & Sitabkhan, Y. (2014). Early grade mathematics assessment (EGMA) toolkit. https://shared.rti.org/content/early-grade-mathematics-assessment-egma-toolkit

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://doi.org/10.32614/r.manuals

Roesch, S., & Moeller, K. (2015). Considering digits in a current model of numerical development. *Frontiers in Human Neuroscience*, *8*, 1062. https://doi.org/10.3389/fnhum.2014.01062

Saalbach, H., Eckstein, D., Andri, N., Hobi, R., & Grabner, R. H. (2013). When language of instruction and language of application differ: Cognitive costs of bilingual mathematics learning. *Learning & Instruction*, *26*, 36–44. https://doi.org/10.1016/j.learninstruc.2013.01.002

Sells, L. (1973). High school mathematics as the critical filter in the job market. Developing Opportunities for Minorities in Graduate Education: Proceedings of the Conference on Minority Graduate Education at the University of California, Berkeley, May, 1973, pp. 39–47.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Researchand Perspectives*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology General*, *116*(3), 250–264. https://doi.org/10.1037//0096-3445.116.3.250

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology General*, *117*(3), 258–275. https://doi.org/10.1037//0096-3445.117.3.258

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 227–293). Erlbaum.

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*, *23*(8), 1–22. https://doi.org/10.1007/s10763-025-10578-3

Sonnleitner, P., Bernard, S., Michels, M., Inostroza-Fernandez, P., Keller, U., Gierl, M., Gierl, M., & Hornung, C. (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*, *38*(2), 95–117. https://doi.org/10.1080/08957347.2025.2563889

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. [Students' longitudinal competencies – Development of German Reading Comprehension and Mathematics in Luxembourg between 3rd and 9th Grade]. Esch-sur-Alzette, Luxembourg: LUCET, Universität Luxemburg, SCRIPT.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45–88. https://doi.org/10.1016/s0010-0277(00)00108-6

Suárez-Pellicioni, M., Núñez-Peña, M. I., & Colomé, A. (2015). Math anxiety: A review on its cognitive consequences, psychophysiological correlates and brain bases. *Cognitive Affective Behavioral Neuroscience*, *16*(1), 3–22. https://doi.org/10.3758/s13415-015-0370-7

Tatsuoka, C., Clements, D. H., Sarama, J., Izsak, A., Orril, C. H., de la Torre, J., & Khasanova, E. (2016). Developing workable attributes for psychometric models based on the q-matrix. *Journal of Research on Mathematics Education*, *15*, 73–96. https://www.jstor.org/stable/26858791

Vasilyeva, M., Laski, E. V., Casey, B. M., Lu, L., Wang, M., & Cho, H. Y. (2023). Spatial-numerical magnitude estimation mediates early sex differences in the use of advanced arithmetic strategies. *Journal of Intelligence*, *11*(5), 97. https://doi.org/10.3390/jintelligence11050097

von Aster, M. G. (2005). Wie kommen zahlen in den Kopf? Ein Modell der normalen und abweichenden Entwicklung zahlenverarbeitender Hirnfunktionen. In M. G. von Aster & J. H. Lorenz (Eds.), *Rechenstörungen bei Kindern. Neurowissenschaft, Psychologie, Pädagogik* (pp. 13–33). Vandenhoeck & Ruprecht.

Von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine and Child Neurology*, *49*(11), 868–873. https://doi.org/10.1111/j.1469-8749.2007.00868.x

Von Aster, M., Schweiter, M., & Weinhold Zukauf, M. (2007). Rechenstörungen bei Kindern.: Vorläufer, Prävalenz und psychische Symptome. *Zeitschrift Fur Entwicklungspsychologie Und Padagogische Psychologie*, *39*(2), 85–96. https://doi.org/10.1026/0049-8637.39.2.85

Von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *Timss, 2023 technical report (methods and procedures)*. TIMSS & PIRLS International Study Center. https://timss2023.org/methods

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*, 352–360. https://doi.org/10.3102/0013189x14553660

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers. Theory into practice*. Springer. https://doi.org/10.1007/978-981-10-3302-5_5

Wu, X., Wu, R., Chang, H. H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology*, *11*, 2230. https://doi.org/10.3389/fpsyg.2020.02230

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS, 2007 fourth grade mathematics assessment. *PLOS ONE*, *13*(2), e0188691. https://doi.org/10.1371/journal.pone.0188691

Zhang, Y., Zhao, B., Jian, M., & Wu, X. (2025). Cognitive diagnostic analysis of mathematics key competencies based on PISA data. *PLOS ONE*, *20*, e0315539. https://doi.org/10.1371/journal.pone.0315539