# WHICH ITEM CHARACTERISTICS DRIVE THE DIFFICULTY OF MENTAL ADDITION AND SUBTRACTION TASKS

Daniel Sommerhoff

IPN - Leibniz Institute for Science and Mathematics Education, Kiel

*Mental addition and subtraction are key learning goals in primary education, yet they remain challenging for learners. Unfortunately, there is limited quantitative evidence on characteristics that influence the difficulty of these tasks (DGICs). Using a dataset with 1,085,842 task performances from 7,307 students across 3,434 tasks, this study quantitatively explores more than 10 DGICs. Results (i) show that all proposed DGICs significantly predict item difficulty, (ii) quantify main and interaction effects, and (iii) underline that the DGICs explain over 50% of the variance in student performance, even without considering students' backgrounds or solution strategies. Overall, this study provides important insights into the difficulty of mental addition and subtraction tasks, informing task design and guiding future research in mathematics education.*

## INTROCUCTION

Arithmetic skills are among the key learning goals in primary education (KMK, 2004; NCTM, 2000). Of the four basic operations, addition and subtraction are especially important because they lay the foundation for later learning of multiplication and division. Research in mathematics education and large-scale educational monitoring studies have repeatedly shown that students struggle with mental calculation in addition and subtraction, and their performance in these areas often falls short of expectations (Hickendorff et al., 2019).

The empirical difficulty of such tasks can be attributed to *task characteristics* (e.g., addition versus subtraction, problem size, crossing a ten), to *student characteristics* (e.g., cognitive abilities, effort, prior knowledge, computational strategies), or to the interaction between the two (e.g., flexibility or adaptivity with which each strategy is chosen for a specific task). Within mathematics education research, solution strategies are often emphasized as a key factor for success (Hickendorff et al., 2019). Likewise, research in mathematical cognition has built a comprehensive body of work on (multi-digit) number processing and its cognitive mechanisms (Nuerk et al., 2015). In contrast, there has been relatively limited research that systematically investigates how multiple task characteristics in addition and subtraction tasks interact and impact on performance. Existing studies are frequently qualitative or based on a small number of tasks, a small number of students (or responses), or both (Heirdsfield & Cooper, 2004).

Using a large dataset from an online learning environment, this study replicates and expands these findings on a stronger empirical basis. It examines which (combination of) task characteristics influence the difficulty of mental addition and subtraction tasks, and the extent to which students' scores can be predicted based on these characteristics.



2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 1379-1386. PME.

## THEORETICAL BACKGROUND

### Mental Computation: Addition and Subtraction

Mental computation typically refers to the ability to perform basic arithmetic operations (e.g., addition, subtraction) in primary education, accurately and quickly, without using any external resources such as paper, pencil, or calculators (Varol & Farran, 2007). It is sometimes framed as arithmetic fluency (Vanbinst et al., 2015) and typically assessed using short, timed arithmetic tasks of the form a±b = __.

Mental computation is included in primary education curricula worldwide because it is highly valuable in everyday life and has been shown to be essential for higher-order mathematical thinking (Hickendorff et al., 2019) and later mathematics learning.

Among the different operations, addition is considered fundamental because subtraction, multiplication, and division can be (re-)interpreted based on repeated addition. Therefore, mental addition skills are considered particularly important.

However, research findings suggest that many students struggle with mental computation throughout and beyond primary school (Hickendorff et al., 2019). Prior studies show that some students' performance in mental addition and subtraction remains low even in higher grades, underscoring the need for further research insights. Poor performance in this area during primary school is largely attributable to task characteristics that increase difficulty (e.g., Benz, 2005) and to the use of (ineffective) solving strategies (Beishuizen et al., 1997; Heirdsfield & Cooper, 2004).

### Difficulty Generating Item Characteristics

Mental calculation tasks in general, especially addition and subtraction tasks within the number domain up to 100, can be classified based on specific Difficulty Generating Item Characteristics (DGICs) (see Anderson et al., 2022). DGICs refer to task characteristics that can differ between tasks and (are assumed to) influence the task's difficulty. For instance, the operation to be performed is considered a DGIC, as research shows that subtraction is generally more difficult than addition. Likewise, the so-called problem-size effect (Zbrodoff & Logan, 2005) highlights that the magnitude of the numbers involved in the task likely affects its difficulty, with larger numbers resulting in a more challenging task (e.g., 2+3 = 5 versus 52+3 = 55).

In their analysis, Anderson et al. (2022) examined the operation, the number of digits of the second number, and the necessity to cross ten as DGICs. They demonstrated that each of these factors represented DGICs that significantly affected item difficulty, that crossing ten was the most important DGIC, and that including three DGICs (without interactions) already accounts for approximately 20% of the variance in item difficulty.

A (non-exhaustive) list of DGICs proposed for addition and subtraction tasks in the number domain of 100 is provided in Table 1. Unfortunately, there is no generally agreed-upon list of DGICs for addition and subtraction tasks, nor is there enough empirical evidence to determine their relative importance for item difficulty. This also

applies because single or a few DGICs often fail to capture item difficulty. For example, the item 35+24—based on the magnitude of the numbers—is likely more difficult than 5+4. However, 30+20—which based only on the DGIC magnitude should be comparable in difficulty to 35+24—is probably much easier since it essentially involves calculating 3+2. Therefore, considering more DGICs and their interactions is necessary. Finally, some DGICs may not be independent; for instance, tasks that involve only multiples of ten typically involve only two-digit numbers (or are trivial).

Table 1: Overview of possible DGICs for addition and subtraction tasks up to 100.

| DGIC | Description | Exemplary tasks |
|---|---|---|
| Operation | The operation that must be performed; addition (0), subtraction (1) | (0) 12+13 = _ vs. (1) 45-12 = _ |
| Crossing ten | The task requires to cross the ten boundary (i.e., borrowing/carrying needed); (0) not necessary, (1) necessary | (0) 13+24 = _ vs. (1) 7+9 = _ |
| Addition to ten | The task requires an addition to ten; (0) not necessary, (1) necessary | (0) 12+13 = _ vs. (1) 22+18 = _ |
| Magnitude first number | The magnitude of the first number in steps of ten; (0) 0-9, (1) 10-19, (2) 20-29, ..., (9) 90-99, (10) 100 | (0) 2+12 = _ vs. (4) 47+4 = _ |
| Magnitude second number | The magnitude of the second number in steps of ten; (0) 0-9, (1) 10-19, (2) 20-29, ..., (9) 90-99, (10) 100 | (0) 12+4 = _ vs. (2) 78-27 = _ |
| Magnitude result | The magnitude of the result in steps of ten; (0) 0-9, (1) 10-19, (2) 20-29, ..., (9) 90-99, (10) 100 | (3) a+b = 35 vs. (7) c-d = 72 |
| Multiples of ten only | The numbers of the task are only multiples of ten; (0) no, (1) yes | (0) 52+30 = _ vs. (1) 50+30 = _ |
| Zero term | The task includes a zero term; (0) no, (1) yes | (0) 2+38 = _ vs. (1) 0+38 = _ |
| Zero result | The result of the task is zero; (0) no, (1) yes | (0) 29-17 = _ vs. (1) 19-19 = _ |
| Duplication | The task represents a duplication task; (0) no, (1) yes | (0) 12+39 = _ vs. (1) 24+24 = _ |
| Halving | The task represents a halving task; (0) no, (1) yes | (0) 12+35 = _ vs. (1) 48-24 = _ |

*Note*: The codes appended to each DGICs' description match the codes used in the analyses of this study. The exemplary tasks systematically vary regarding the respective DGIC, but usually also regarding other DGICs.

## RESEARCH QUESTIONS

While the fields of mathematics education and mathematical cognition have produced extensive research on addition and subtraction over the past decades, there is currently limited quantitative evidence for a systematic overview of DGICs of addition and subtraction tasks in the number domain up to 100. To fill this research gap, we focus on the following research question:

RQ1: What are the most important item characteristics that drive the difficulty of addition and subtraction tasks up to 100?

RQ2: How do these item characteristics interact?

RQ3: How much variance in learners' scores can be explained by the DGICs?

## METHOD

### Data Origin

To address the research questions, data from the German online learning platform www.levumi.de were used. The platform provides learning progress monitoring tests for teachers (see Deno et al., 2009; Fuchs et al., 2025), allowing teachers to track students' learning development by administering parallel mathematics tests repeatedly. The platform offers numerous tests, some of which focus on mental addition and subtraction within the number domain up to 100. Each test comprises a set of tasks randomly assigned to students to avoid excessive repeated testing effects.

### Data and Sample

The data retrieved from the platform consist of 1,085,842 task performances by 7,307 students across 946 classes, based on 3,434 unique tasks within the number domain up to 100 (i.e., with results ranging from 0 to 100) of the form a±b = __. These tasks were administered between September 2019 and October 2025. The frequency of task performance per task ranged from 8 to 6020 ($M$ = 316.20; $SD$ = 596.97), and average scores ranged from 0 to 0.94 ($M$ = 0.60; $SD$ = 0.16). Data include a student ID and class ID, allowing to control for respective dependencies, along with details about the specific task administered, the solution entered, the score (0/1), the response time (in ms), and the date of administration. For some students, information regarding gender, age, and special needs was voluntarily provided.

All tests included in this study focus on mental addition and subtraction up to 100, typically administered in second and third grades in Germany (ages approx. 7-9).

### Data Preprocessing

Prior to the analyses, all solutions that showed meaningless processing were removed. Specifically, all task performances with response times less than 1000ms were excluded, as this is below the typical response time for arithmetic fact retrieval. Additionally, all task performances with solutions containing more than four digits were removed (e.g., „999999“, “11111”), since these can be seen as bogus numbers

entered by students to move on to the next task. Finally, all task performances with missing data in relevant categories (see data analysis) were excluded.

Based on this preprocessing, 876,233 task performances from 3,434 tasks and 6,217 students across 794 classes were included. To ensure the integrity of the data cleaning process, all analyses shown below were also conducted on the entire dataset.

### Data Analysis

To address RQ1 and RQ2, generalized linear mixed models (GLMMs) were used, with each task's score (dichotomous, 0/1) as the dependent variable and the DGICs from Table 1 along with gender as fixed factors. Random intercepts for student ID and class ID were included to account for dependencies in the data. The glmmTMB package and R version 4.5.1 were used for the analyses.

For RQ1, an additive base model was calculated. For RQ2, multiple models, including various interactions among the factors, were calculated. To evaluate the overall model fit, models were compared using fit indices and likelihood ratio tests.

To estimate the amount of explained variance in students' scores (RQ3), conditional (random effects) and marginal (random & fixed effects) $R^2_{Nakagawa}$ were calculated.

## RESULTS

The additive base model (RQ1) is shown in Table 2. The estimates indicate the change in item difficulty in logits when the DGIC is present compared to the reference category with level 0 for all DGICs (i.e., addition tasks with numbers and results between 1-9, which are not duplication tasks). On average, reference items have an item difficulty of 2.99 logits, corresponding to a solution rate of 95% among female students. For example, if an item also involves crossing ten (automatically implying a result magnitude of 1, i.e., results 11 to 19), the estimate adjusts to 2.99-0.97-0.33 = 1.69 logits, or a solution rate of 84%.

As expected, all DGICs significantly affected item difficulty. Specifically, the DGICs *operation*, *crossing ten*, and *addition to ten* all increase item difficulty, with *crossing ten* having the greatest impact. In terms of magnitude, item difficulty generally increases with increasing *magnitude of the first number*, reaching a plateau after around 60. Overall, increasing the *magnitude of the second number* (*est* < -.9) results in more difficult items compared to *first number's magnitude* (*est* > -.9). While a *magnitude of the result* larger than 9 generally leads to more difficult items (estimates approx. -.3 to -.4), no clear pattern emerges between the different DGIC levels.

The DGICs for special tasks show that tasks involving *multiples of ten* are less difficult. Similar patterns appear for tasks with a *zero result*, *duplication* tasks, and *halving* tasks.

*Gender*, as the only person-related variable in the analyses, which served only as a control variable, reflects that boys, on average, demonstrated significantly better skills, but only with a comparatively small effect size.

Table 2: Results from the generalized linear mixed model (RQ1).

| Factors | Estimate | *SE* | *z* | *p* |
|---|---|---|---|---|
| Intercept | 2.99 | 0.04 | 68.36 | < .001 |
| Operation | -0.45 | 0.01 | -47.64 | < .001 |
| Crossing ten | -0.97 | 0.01 | -123.39 | < .001 |
| Addition to ten | -0.14 | 0.01 | -12.74 | < .001 |
| Magnitude first number 1 | -0.38 | 0.01 | -27.00 | < .001 |
| Magnitude first number 2 | -0.57 | 0.02 | -29.14 | < .001 |
| Magnitude first number 3 | -0.70 | 0.02 | -31.50 | < .001 |
| Magnitude first number 4 | -0.76 | 0.02 | -33.82 | < .001 |
| Magnitude first number 5 | -0.83 | 0.02 | -35.11 | < .001 |
| Magnitude first number 6 | -0.87 | 0.02 | -35.10 | < .001 |
| Magnitude first number 7 | -0.85 | 0.02 | -34.31 | < .001 |
| Magnitude first number 8 | -0.84 | 0.03 | -31.87 | < .001 |
| Magnitude first number 9 | -0.81 | 0.03 | -27.89 | < .001 |
| Magnitude second number 1 | -0.90 | 0.01 | -73.85 | < .001 |
| Magnitude second number 2 | -1.10 | 0.02 | -64.15 | < .001 |
| Magnitude second number 3 | -1.17 | 0.02 | -64.49 | < .001 |
| Magnitude second number 4 | -1.12 | 0.03 | -43.51 | < .001 |
| Magnitude second number 5 | -1.19 | 0.03 | -45.35 | < .001 |
| Magnitude second number 6 | -1.33 | 0.03 | -41.10 | < .001 |
| Magnitude second number 7 | -1.33 | 0.05 | -27.62 | < .001 |
| Magnitude second number 8 | -1.34 | 0.05 | -26.71 | < .001 |
| Magnitude second number 9 | -1.01 | 0.19 | -5.21 | < .001 |
| Magnitude result 1 | -0.33 | 0.01 | -25.38 | < .001 |
| Magnitude result 2 | -0.33 | 0.02 | -18.30 | < .001 |
| Magnitude result 3 | -0.43 | 0.02 | -20.29 | < .001 |
| Magnitude result 4 | -0.37 | 0.02 | -16.50 | < .001 |
| Magnitude result 5 | -0.35 | 0.02 | -15.57 | < .001 |
| Magnitude result 6 | -0.40 | 0.02 | -17.11 | < .001 |
| Magnitude result 7 | -0.31 | 0.02 | -12.91 | < .001 |
| Magnitude result 8 | -0.37 | 0.03 | -14.67 | < .001 |
| Magnitude result 9 | -0.29 | 0.03 | -10.96 | < .001 |
| Magnitude result 10 | -0.28 | 0.04 | -6.72 | < .001 |
| Multiples of ten only | 1.22 | 0.03 | 43.70 | < .001 |
| Zero term | -0.05 | 0.02 | -2.71 | .007 |
| Zero result | 0.44 | 0.03 | 14.65 | < .001 |
| Duplication | 0.48 | 0.03 | 18.07 | < .001 |
| Halving | 0.14 | 0.03 | 4.54 | < .001 |
| Gender female | 0.27 | 0.04 | 7.24 | < .001 |

*Note*: Estimates are presented in logits. The numbers for the magnitude DGICs refer to the respective level, i.e. “magnitude second number 4” to a second number between 40 and 49.

Overall, the additive model (AIC = 639274.9) shows a good model fit and is significantly better compared to a null model with only random effects for student and class (LRT: $\chi^2(37) = 46607$; $p < .001$).

Regarding RQ2, multiple models have been estimated, which cannot be fully presented here. Results show several significant interaction effects, such as *operation***crossing ten*, highlighting that borrowing in subtraction is slightly easier than carrying in addition. Including these interactions generally leads to an improved model fit in terms of fit indices (e.g., model with the interaction *operation***crossing ten*: AIC = 639087.7) and a significantly better model fit compared to the additive model (e.g., model with the interaction *operation***crossing ten*: LRT: $\chi^2(1) = 189.23$; $p < .001$).

Finally, regarding RQ3, the calculated models show a high explained variance of $R^2_{Nakagawa,conditional} > 50.2\%$, with 50.2% for the additive model. The random intercepts for learner and class explain $R^2_{Nakagawa,marginal} \approx 15\%$ of the variance

## DISCUSSION

Gaining insights from a large, authentic dataset into difficulty generating item characteristics of addition and subtraction tasks within the number domain up to 100 is essential because: i) it allows to confirm qualitative hypotheses about DGICs from previous research; ii) it provides further understanding of the problems and issues students face (e.g., Benz, 2005; Hickendorff et al., 2019); and iii) it may enable better planning and allocation of tasks in learning environments. For example, based solely on the additive base model, new insights can be gained into the power-effect and the difficulty of specific item formats. For example, it is quite interesting that an addition to ten (e.g., 6+4 = __) is much (approx. 0.83 logits) easier than crossing the ten (e.g., 7+4 = __), even though both tasks require a carry operation. Moreover, the explained variance by the models is above 50%, solely based on information on DGICs and gender, without attending to further information, for example, regarding the strategies used by the learners. This highlights the importance of research on the DGICs and sound empirical evidence regarding their specific impact on difficulty, as task difficulty can be estimated quite well from the DGICs alone.

However, the results presented here, which focus on the additive base model, should be interpreted with caution because the DGICs are not independent of one another. For example, if a task involves multiples of ten only, which reduces item difficulty (est.: +1.22 logits), it typically also involves higher levels of the magnitude of the first and second numbers as well as the result. This, in turn, increases item difficulty (combined approx. -2.3 logits, depending on the actual magnitude levels), resulting in an overall increase in difficulty of approximately -1.1 logits. It is thus crucial to consider interactions and make planned contrasts between different types of items. Additionally, the data must be handled carefully because there is little background information available, such as learners' ages, grade levels, or the context in which the test is taken.

Overall, the analyses above offer strong empirical insights into the complex landscape of tasks and DGICs. They complement prior qualitative findings from mathematics education (Hickendorff et al., 2019) as well as precise yet narrow experimental insights from mathematical cognition (Nuerk et al., 2015). Additionally, the analyses highlight

the importance of DGICs, complementing work on learners' strategies and on more complex word problems that pose additional demands.

## References

Anderson, S., Sommerhoff, D., Schurig, M., Ufer, S., & Gebhardt, M. (2022). Developing learning progress monitoring tests using difficulty-generating item characteristics: An example for basic arithmetic operations in primary schools. *Journal for educational research online*, *2022*(1), 122-146.

Beishuizen, M., Van Putten, C. M., & Van Mulken, F. (1997). Mental arithmetic and strategy use with indirect number problems up to one hundred. *Learning and Instruction*, *7*(1), 87-106.

Benz, C. (2005). *Erfolgsquoten, Rechenmethoden, Lösungswege und Fehler von Schülerinnen und Schülern bei Aufgaben zur Addition und Subtraktion im Zahlenraum bis 100 [Students' success rates, calculation methods, solutions and mistakes in addition and subtraction tasks in the range up to 100]*. Franzbecker.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools*, *46*(1), 44-55.

Fuchs, A., Radkowitsch, A., & Sommerhoff, D. (2025). Using learning progress monitoring to promote academic performance? A meta-analysis of the effectiveness. *Educational Research Review*, *46*, 100648.

Heirdsfield, A. M., & Cooper, T. J. (2004). Factors affecting the process of proficient mental addition and subtraction: case studies of flexible and inflexible computers. *The Journal of Mathematical Behavior*, *23*(4), 443-463.

Hickendorff, M., Torbeyns, J., & Verschaffel, L. (2019). Multi-digit Addition, Subtraction, Multiplication, and Division Strategies. In *International Handbook of Mathematical Learning Difficulties* (pp. 543-560).

Kultusministerkonferenz (KMK). (2004). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. (Beschlüsse der Kultusministerkonferenz).

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. National Council of Teachers of Mathematics.

Nuerk, H.-C., Moeller, K., & Willmes, K. (2015). Multi-digit Number Processing: overview, conceptual clarifications, and language influences. In R. Cohen Kadosh & A. Dowker (Eds.), *The Oxford Handbook of Numerical Cognition* (pp. 106-139). OUP.

Vanbinst, K., Ceulemans, E., Ghesquiere, P., & De Smedt, B. (2015). Profiles of children's arithmetic fact development: a model-based clustering approach. *J Exp Child Psychol*, *133*, 29-46.

Varol, F., & Farran, D. (2007). Elementary School Students' Mental Computation Proficiencies. *Early Childhood Education Journal*, *35*(1), 89-94.

Zbrodoff, N. J., & Logan, G. D. (2005). What everyone finds: The problem-size effect. In *The handbook of mathematical cognition* (pp. 331-345). Psychology Press.