MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2026, 22(1), em2755
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/17564

Factors affecting the difficulty of comparison-type word problems in primary mathematics

Eszter Kónya [1,2*], Balázs Vértessy [2,3]

[1] University of Debrecen, Debrecen, HUNGARY
[2] MTA-DE University of Early Science Learning Research Group, Debrecen, HUNGARY
[3] Nemeskürty István Faculty of Teacher Training, Ludovika University of Public Service, Budapest, HUNGARY

Received 25 February 2025 ▪ Accepted 28 September 2025

Abstract

This study investigates factors influencing the difficulty of compare word problems (WPs), a crucial theme for early mathematical development, often linked to quantitative observations. Four factors are examined: the required operation (addition/subtraction vs. multiplication/division), the lexical consistency (match between problem-wording and required operation), the role of the unknown (compared set, comparative relation, or reference set), and order-consistency (the order of numbers in the text is consistent or inconsistent with them in the operation sentence). A written survey containing single-operation WPs was administered to 651 students of grade 3 and grade 4 from Eastern Hungary. The proportion of correct responses measured the item's difficulty. Results confirmed the strong influence of required operation and lexical consistency in line with prior research. Furthermore, a similar difficulty order based on the unknown type was observed in additive and multiplicative compare problems. However, while influential, order-consistency proved less decisive than the other factors.

Keywords: word problem, comparison problem, linguistic structure of the problem, problem difficulty

INTRODUCTION

Arithmetic operations are presented through interpreting different everyday situations in primary school. In Hungary, additive operations (addition and subtraction) are introduced in grade 1 and multiplicative (multiplication, division) in grade 2. Each operation that students learn is linked to various everyday situations; more precisely, it is abstracted from them. Finding an arithmetic model for simple word problems (WPs) is a fundamental expectation in the early grades. It is also essential for a science, technology, engineering, and mathematics (STEM) approach. The Hungarian core curriculum for grade 1-grade 4 (Hungarian Government, 2020) requires the integration of science into other subjects, such as mathematics. Therefore, mathematics instruction should contribute to the development of STEM competencies by fostering quantitative reasoning and linking mathematical structures to real-world and scientific contexts. As science is taught only one lesson per week in grade 3 and grade 4 (and none in grade 1 and grade 2), mathematics lessons offer a crucial opportunity to cultivate STEM practices, including observing and measuring quantities, comparing magnitudes, and constructing models. Comparison-type WPs are particularly well suited for this purpose: they help students identify relevant quantities of everyday or scientific situations, recognize the quantitative relationships between them, and express these relationships in mathematical form.

Prior research has consistently shown that WPs are challenging for students well beyond the early grades (Nunes et al., 2016; Riley et al., 1983). There is broad agreement that problem difficulty is influenced by multiple factors, including the operation required and the presence of lexical consistency or inconsistency in the text (Lewis & Mayer, 1987; Selter, 2000). Findings also converge on the importance of the role of the unknown: reference-set problems are systematically harder than those with a compared set (CS) (Carpenter et al., 1981; Daroczy et al., 2015). Where literature diverges is in the

✉ eszter.konya@science.unideb.hu **(*Correspondence)** ✉ vertessy.balazs@uni-nke.hu

Contribution to the literature

- There is a rich literature on the factors that influence the difficulty of compare WPs in primary grades. Our study confirms previous findings on required operation and lexical consistency.
- The order of difficulty by unknown type was as follows: CS, comparison relation, and reference set (RS). While this hierarchy is well established for additive WPs, our results extend it to multiplicative problems as well.
- In contrast to earlier findings, our experience is that order-consistent problems can be even more difficult than order-inconsistent ones.

effect of order-consistency: some studies suggest that problems where numbers appear in the same order in the text as in the operation are easier, while others find negligible or even reversed effects (Daroczy et al., 2015).

Despite this extensive work, less is known about how these factors operate jointly, particularly in the case of comparison-type WPs at the primary level. The present study addresses this gap by systematically examining the combined effects of different factors in 12 carefully designed one-step comparison-type WPs. By analyzing how particular task characteristics influence the selection of the appropriate arithmetic model, the aim is to highlight the interpretative difficulties inherent in the wording of the tasks. This knowledge may be essential for designing future instructional interventions that strengthen quantitative reasoning and contribute to the integration of STEM practices in primary mathematics education.

THEORETICAL FRAMEWORK

A mathematical task is considered a WP when the answer results from a mathematical calculation; however, the data needed to do this are given in text rather than numbers (Selter, 2000). Specifically, arithmetic WPs can be solved by simple arithmetic operations, often by mental calculation. The difficulty can be in deciding which operation to perform on which data, based on the text.

In their paper, Daroczy et al. (2015) reviewed, among other things, factors that influence the difficulty of WPs. Three main categories were distinguished: linguistic factors, numerical factors, and interaction between linguistic and numerical factors.

Linguistic complexity manifests in structural and semantic factors. Structural factors include the number of words and sentences, unknown words, grammatically complex words, etc. The order of the numbers in the text is also considered a structural factor. Its effect depends on whether it differs from the order of numbers used when performing the necessary operation (if the two orders are the same, a better result is expected). In what follows, we refer to this factor as the order-consistency.

Researchers in the 1980s and earlier also found that semantic factors can cause significant differences between problems (Riley et al., 1983). They turned their attention to the semantic structure of additive WPs.

The categorization of single-operation additive WPs is well-known in the literature and originated from a summative study by Riley et al. (1983). They highlighted different types of WPs (change, combine, and compare) that influence the difficulty. Other researchers later adopted these categories (Cummins, 1991; de Corte & Verschaffel, 1987; Nunes et al., 2016). Change problems involve transformation, i.e., the number of particular objects increases or decreases (Nunes et al., 2016). Three cases can be distinguished depending on whether the initial quantity, the number of changes, or the resulting quantity is unknown. Combine problems involve the composition of two quantities. Using set operations, this type of problem can be described as being about the number of elements of the union of two disjunct sets (addition) or the number of elements of the difference set of a set and its subset (subtraction) (Herendiné Kónya, 2013). Compare problems involve comparison relations. The question can be about the difference (how much more/less a quantity is than the other), the compared quantity, or the referent quantity. The combine and compare problems are static, while the change problems have a dynamic feature.

Most of the literature is concerned with the semantic structure of additive and, to a lesser extent, with multiplicative WPs. However, Greer (1992) distinguished four categories. This is the starting point for multiplicative categorization, which applies to multiplying whole numbers. These categories are as follows:

1. Equivalent groups (e.g., 2 tables, each with 4 children), a repeated addition, is considered a primitive multiplication model. It fits the additive combine problem well and works with a positive integer multiplier.
2. Multiplicative comparison (e.g., 3 times as many boys as girls).
3. Rectangular arrays (e.g., 3 rows, each with 4 children).
4. Cartesian product (e.g., the number of possible boy-girl pairs); it is related to combinatorial problems.

de Corte et al. (1988) refer to Greer's (1987) earlier classification scheme and state that multiplicative problems can be divided into two basic groups: symmetric and asymmetric. In asymmetric multiplication situations, the multiplier and the multiplicand can be distinguished. WPs belonging to the category of equivalent groups describe such asymmetric situations. In the example (2 tables, each with 4 children), the multiplier is 2, and the multiplicand is 4. This multiplication problem can be turned into a division problem in two ways. First, the task leads to a quotative division (8 children must sit in the dining room, where 4 chairs belong to each table. How many tables do they need?). Second, a partitive division problem can also be constructed (8 children are divided equally among 2 tables. How many children sit around one table?). Further asymmetric multiplicative problems involve the iteration of measure, change of scale, rate, and measure conversion. The change of scale and rate problems can be considered multiplicative compared to problems. A multiplicative problem is symmetric when the distinction between the multiplier and the multiplicand is unnecessary because the two numbers' roles are the same. Rectangular arrays, Cartesian products, and area problems belong to this category. Another categorization also appeared in the literature (Nunes et al., 2016), but in the present research, Greer's (1987) categories, supplemented by the appropriate division categories, are the most helpful.

The current analysis concentrates mainly on semantic structures but also considers certain structural factors. The numerical complexity of a WP may depend on the type of numbers presented (multi-digit numbers, whole numbers, or fractions), the number of operations required, the number of solving steps, the solving strategy, etc. (Daroczy et al., 2015; Pongsakdi et al., 2020). Among the numerical factors, this study focuses on the type of operation. The other aspects are the simplest: whole numbers in the hundreds and tasks that can be solved with one operation.

The interaction between linguistic and numerical factors influences the solution of arithmetic WPs. The lexical consistency effect (Daroczy et al., 2015) belongs to this category. A lexically consistent/inconsistent WP is a problem in which the relational terms and linguistic cues in the text (e.g., *more, less, altogether,* and *remains*) are consistent/conflict with the arithmetic operation required for its solution. In line with the consistency hypothesis (Lewis & Mayer, 1987), in the case of lexically inconsistent problems, solvers must reinterpret or mentally invert the linguistic relation to construct the correct arithmetic model, which typically increases the likelihood of errors. An example of lexically inconsistent WP: "If I spent 120 HUF of my money, I would only have 310 HUF left. How much money do I have now?" (Herendiné Kónya, 2013, p. 169). The words *spent* and *left* suggest the operation of subtraction, but the answer is 310 + 120 = 430. It is worth noting that the Hungarian version of the text has a very similar structure and contains the same keywords (spent = elköltenék, left = maradna), so the aspect of lexical consistency is not affected by linguistic differences.

Riley et al. (1983) reported a global structural factor, the unknown's position in the text. If the first or second term of the operation in the sentence is missing (i.e., □ + b = c or a + □ = c), the solution is more challenging than if the result of the operation (a + b = □) is the question. This situation often happens when the WP has inconsistent language. In the above problem, the first subtraction term is missing (□ - 120 = 310). Since this factor is highly dependent on lexical consistency, it will not be addressed separately in this study.

Daroczy et al. (2015) have also identified other factors related to the interaction of linguistic and numerical issues that influence the success of WP solutions. Division problems often involve functionally related objects (e.g., tulips and vases), while additional problems involve objects in the same category. A structural correspondence between the semantic and mathematical content contributes to success (Páchová & Vondrová, 2021). The multiplicative problems in this study are, with one exception, compare WPs, so this factor does not appear because objects of the same category are being compared. The general experience is that numerical or linguistic content irrelevant to the solution is a burden. This is the case for tasks involving redundant data or information, as the first step is to filter out content relevant to the solution. However, this survey covers problems where this aspect is not present.

Results of Experimental Research Related to Additive Word Problems

Riley et al. (1983) were among the first conducted surveys to determine which additive WPs are more complicated than others for first graders and older students. Research in this domain has led to convergent conclusions across studies in different countries. We list some results related to the present study.

1. It is well-known from the literature that compare problems are more difficult than change and combine problems for first graders (Carpenter et al., 1981; Nunes et al., 2016; Riley et al., 1983). Also, for subtraction, the order of difficulty is the same. Furthermore, the papers referred to the fine structure of the compare WPs. The comparison situation involves three quantities: the RS, the CS, and the comparative relation (CR). The following example explains the meaning of these notions: "Alex has 13 books. He has 5 more books than Camilla. How many books does Camilla have? (Nunes et al., 2016, p. 17)". The RS is Alex's books (13), the CR is *5 more*, and the CS is Camilla's (unknown). Citing several studies, Nunes et al.

(2016) pointed out that, among the compare problems, the most difficult were those in which the unknown is the RS, while the easiest were those in which the CS is.

2. For beginners, the position of the unknown in the text affects the difficulty level as follows: in change problems, when the start and change amounts are given, and they are asked for the result, it is the easiest (Riley et al., 1983). de Corte and Verschaffel (1987) also stated that determining the required operation is affected by the position of the unknown in the text. Nunes et al. (2016) confirmed that this difficulty remains in higher grades and other additive WPs.
3. Lexical inconsistency concerning the required mathematical operation is an essential determinant of task difficulty. Inconsistent language results in a high error rate and extended response time (Hegarty et al., 1992). Verschaffel's (1994) research on one-step compare WPs with fifth graders provided strong empirical evidence for the consistency hypothesis of Lewis and Mayer (1987). Similar results were obtained from an eye-tracking study of 10-11-year-old students conducted by Bartalis et al. (2023).
4. According to de Corte et al. (1990), the choice of the correct operation depends strongly on the type of numbers involved. Results of experimental research related to multiplicative WPs.

Results of Experimental Research Related to Multiplicative Word Problems

5. Equivalent groups problems are more straightforward than Cartesian product problems (de Corte et al., 1988).
6. Students' difficulty in choosing the correct operation depends on the type of numbers in the problem (de Corte et al., 1988). WPs with whole numbers are easier than with decimal fractions. In other words, tasks related to whole numbers are easier than others with the same semantic structure, even if the requirement is selecting the correct operation without solving the problem itself. It was found that the type of the multiplier strongly influences the students' results, while the effect of the type of the multiplicand is much smaller.
7. Multiplication and partitive division problems are equally challenging but more manageable than quotative ones (Greer, 1992).
8. Pape (2003) extended the investigation of the consistency hypothesis for middle school students' problem-solving using both additive and multiplicative compare problems. In multiplication and division compare problems, the wording *n times as many* and *1/n as many* signify the relation. His research indicates a strong impact of fraction-of-a-number relational terms on the success of multiplicative compare problems, independent of the lexical consistency effect (in the case of both inconsistent language multiplication and consistent language division problems). These problems caused considerable difficulties for middle school students, possibly because of the primitive intuitive model of multiplication, i.e., repeated addition.

Students' Solution Strategies

Empirical evidence has convincingly shown that the semantic structure of WPs strongly impacts the difficulty and the strategies that young children apply when solving arithmetic WPs (Pongsakdi et al., 2020).

Mayer and Hegarty (1996) identified two main strategies when primary students solve simple arithmetic WPs. A direct translation strategy means selecting numbers from the problem and then performing arithmetic operations on them. The problem model strategy is used when trying to understand the situation described in the problem and devising a solution plan based on the ensuring representation of the situation. Many unsuccessful problem solvers rely on the direct translation strategy and fail to provide correct answers, mainly when problems include important implicit information, which they should infer based on the situation described in the text. The authors concluded from their research series on the reading of WPs that successful students are more likely than unsuccessful students to use a problem model strategy, while the opposite is confirmed by using a direct translation strategy. Practicing inconsistent language problems provides an excellent opportunity to use the problem model strategy instead of the translation strategy.

The Function of Word Problems in the Teaching Process

Szomjú and Habók (2015) pointed out that WPs in mathematics classes are often used to practice performing the operations that have just been learned (application function instead of arithmetical modelling function of WPs). This way, the practice brings implicit assumptions restricting students' reasoning and willingness to reason. This can result in students neglecting to understand the meaning of the text and thus accepting unreasonable responses that cannot occur in real life.

In contrast, Nunes et al. (2016) argue that quantitative reasoning and arithmetic are different abilities and must be seen as a domain of teaching and learning on their own; students need to learn to reason about relationships between quantities to solve problems, not only about arithmetic. Considering that WPs are texts

Table 1. The number of participants

	School 01	School 02	School 03	School 04	School 05	School 06	School 07	School 08	Total
Grade 3	119	44	22	43	27	11	15	76	357
Grade 4	83	59	13	37	23	8	0	71	294
Total	202	103	35	80	50	19	15	147	651

that typically contain quantitative information and "describe a situation assumed familiar to the reader and pose a quantitative question, an answer to which can be derived by mathematical operations performed on the data provided in the text, or otherwise inferred" (Greer et al., 2002, p. 271), they focus on quantitative reasoning rather than arithmetic in solving WP's.

Considering all these experiences, 12 additive and multiplicative WPs were selected for the current survey mainly compare problems that can be characterized by lexical consistency, unknown type, and order-consistency.

RESEARCH QUESTION

To better understand the fine structure of compare WPs to develop quantitative reasoning later in STEM education, a survey was conducted to map the solutions of the 3rd and 4th graders. The following research question was formulated.

To what extent do the type of operation, lexical consistency, role of the unknown, and order-consistency affect the success rate of third and fourth graders in finding the arithmetic model of compare WPs?

The texts in the WPs are inspired by situations from the students' everyday lives.

METHOD

Sample

Students in grade 3 and grade 4 (ages 9-11) from eight primary schools in Eastern Hungary participated in the survey (**Table 1**). The schools were selected from schools in a large city (School 01, 08), towns (School 02, 04, 05) and villages (School 03, 06, 07) to capture a diverse sample. Although not strictly random, the sample represents a broad range of school contexts and achievement levels. Within schools, intact classes participated, resulting in a total of 651 students (357 in grade 3 and 294 in grade 4). No one from the classes was excluded, except for students who were absent from school on the given day. The participating students, therefore, represented approximately 90% of the schools' 3rd and 4th grade students.

The survey was conducted in regular classroom settings under the supervision of the class teacher, who provided no additional instructions or assistance in completing the test. The paper-pencil tests took thirty minutes. Ethical review and approval were waived for this study because the survey consisted exclusively of solving mathematics WPs during regular classroom instruction. Students' participation did not involve any interventions beyond usual educational practice, and no personal data or opinions were collected. Responses were recorded anonymously, and no identifying information was stored. Written permission to conduct the survey was obtained from the principals of all participating schools.

Word Problems in the Survey

Knowing that numerical and linguistic complexity can add up (Daroczy et al., 2015), 12 simple arithmetic WPs requiring one operation were constructed, using common words, and working with numbers 0-100 [1]. The exact two numbers were used in each task: 3 and 33, because choosing the correct operation strongly depends on the type of the given numbers in the problem (de Corte et al., 1990). According to the Hungarian national core curriculum (Hungarian Government, 2020) and the relevant framework curriculum (The Educational Authority Hungary, 2020), the knowledge needed to solve such problems correctly is part of the requirements for 2nd graders.

Research shows that performance is also affected by the form in which the answer is given: selecting the correct result, or the required operation, or performing the correct calculation (Pape, 2003). Performing the calculation was not requested; only the selection of the required operation was: "How would you calculate it? Write in the appropriate operation sign: 33 □ 3". Therefore, the students were forced to interpret the simple text as a whole and could focus on the arithmetic modelling of simple everyday situations. However, before solving the WPs, students were asked to perform the four operations to ensure they identified the operation result with the correct numbers.

Five factors characterize our WPs:

1. The required operation is additive or multiplicative (hereafter briefly referred to as the operation).
2. Lexical inconsistency appears or not (hereafter briefly referred to as the lexical consistency).
3. The semantic category is compare or other (hereafter briefly referred to as the semantic category).

[1] The design of the problems was inspired by the tasks of Erdész and Kovács (1982, p. 16).

Table 2. The factors of the WPs

WP	Semantic category	Operation	Lexical consistency	Role of unknown	Order-consistency	Answer
P1	Compare	Additive	Yes	CS	Yes	33 + 3
P2	Compare	Additive	No	RS	Yes	33 - 3
P3	Compare	Additive	Yes	CS	No	33 - 3
P4	Change	Additive	Yes	Result	Yes	33 - 3
P5	Compare	Additive	Yes	CR	No	33 - 3
P6	Equivalent groups	Multiplicative	Yes	Result	No	33 × 3
P7	Compare	Multiplicative	No	RS	No	33 ÷ 3
P8	Compare	Additive	No	CR	Yes	33 - 3
P9	Compare	Multiplicative	No	RS	Yes	33 × 3
P10	Compare	Multiplicative	Yes	CS	Yes	33 ÷ 3
P11	Compare	Multiplicative	No	CR	Yes	33 ÷ 3
P12	Compare	Multiplicative	Yes	CR	No	33 ÷ 3

4. The unknown is the CR, the CS or the RS (hereafter briefly referred to as the role of unknown).
5. The order of the numbers in the text, and the operation is the same or different (hereafter briefly referred to as the order-consistency).

The factors are illustrated by two examples.

P1. Boti collected 33 chocolate eggs at Easter, and Ákos collected 3 more eggs than Boti. How many chocolate eggs did Ákos collect? (answer: 33 + 3).

The solution required an additive operation, namely addition. Lexical inconsistency does not appear because the keyword *more* refers to addition. This is a compare problem, where the unknown is the CS, the number of Ákos's eggs. The WP is order-consistent, since the order of numbers in the text (33 and 3) and the operation (33 + 3) are the same.

P9. My brother and I collected seashells during the holiday. My brother collected 33 shells, a third of my shells. How many shells did I collect during the holiday? (answer: 33 × 3).

The solution required a multiplicative operation, namely multiplication. Lexical inconsistency appears because the keyword *third* refers to division, not multiplication. This is a compare problem, where the unknown is the RS, the number of shells collected by the narrator. The WP is order-consistent, since the order of numbers in the text (33 and 3) and the operation (33 × 3) are the same.

Table 2 summarizes the factors of the 12 WPs. Although complete orthogonality was not feasible, the problem design aimed to approximate balanced coverage of the factor space.

The 12 problems appeared in random order on the test sheet.

The 10 WPs were constructed so that some of them differ in only one factor, allowing for targeted comparison. **Table 3** details the extent of variation across the other four factors. The highlighted pairs are distinguished by one factor only (the pairs in yellow vary only in the operation; those in blue in the unknown, and those in green in the order of the numbers).

Table 3. The number of different factors of problem pairs

	P1	P2	P3	P5	P7	P8	P9	P10	P11	P12
P1		2	1	2	4	2	3	1	3	3
P2			3	3	2	1	1	3	2	4
P3				1	3	3	4	2	4	2
P5					3	2	4	3	3	1
P7						3	1	3	2	2
P8							2	3	1	3
P9								2	1	3
P10									2	2
P11										2
P12										

One point was assigned for each correct answer (item), giving 12 points for the survey. This also means that the mean score calculated for a given problem equals the proportion of correct answers in the sample.

Establishing the Difficulty Level of Word Problems

The difficulty level of a given WP is determined based on the classical test theory; namely, the proportion of individuals answering the item correctly is used as the index for the item difficulty (Pongsakdi et al., 2020). This method met the secondary objective of the survey, namely the intention to use the results to design a future teaching experiment in the participating schools.

Data Analysis Method

First, a descriptive statistical analysis of the results is carried out, and then the significance of the differences is examined. Each statistical test is chosen for its suitability to the nature of the data and the research objectives.

First, the overall scores for the test are analyzed, and then use the Mann-Whitney U test is used to assess the performance of independent groups, such as 3rd and 4th graders, on the same type of WPs. This test is suitable because the data are ordinal, as students' performance was measured by the number of correct answers, which can be ranked but may not be normally distributed. It is

Table 4. The success rate for the four operations

Operation	Total		Grade 3		Grade 4		Comparison	
	n	%	n	%	n	%	φ	p
33 + 3	637	97.85	352	98.60	285	96.94	.06	.146
33 - 3	642	98.62	353	98.88	289	98.30	.03	.528
33 × 3	573	88.02	311	87.11	262	89.12	.03	.434
33 ÷ 3	593	91.09	308	86.27	285	96.94	.19	<.001

Table 5. The average performance of participants

Grade 3		Grade 4		Total	
Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
7.66	2.85	9.27	2.14	8.38	2.68

a non-parametric alternative to the independent samples t-test, which is robust to violations of normality and unequal group sizes and allows us to compare median performance levels between groups. The correct answers are also compared task-by-task using Pearson's Chi-square test, which determines the association between categorical variables, such as problem type and students' success rate (mean score). This non-parametric test does not assume a normal distribution, making it ideal for analyzing frequency data. The Wilcoxon signed-rank test compares paired data, such as students' performance on additive versus multiplicative problems or lexically consistent versus inconsistent problems, etc. This test is a non-parametric alternative to the paired samples t-test, making it appropriate for ordinal or non-normally distributed data, allowing for the assessment of whether there were significant differences in performance across problem types within the same group of students. After selecting task pairs that differ in only one factor, the scores are compared using the McNemar test. It is designed to assess differences in proportions of paired binary outcomes, making it suitable for analyzing whether students' correct or incorrect responses differ significantly across two conditions. The McNemar test complements the Wilcoxon test by providing a focused analysis of categorical response changes rather than ordinal rankings. To control the increased risk of type I error due to multiple comparisons, all results are tested at $\alpha = .01$.

All analyses were carried out using SPSS (30.0.0).

RESULTS

The Average Performance of Students

Before analyzing the performance in detail, it was checked whether students could complete the four required operations (**Table 4**). The percentages in the **Table 4** indicate the percentage of students who can solve the given operation correctly. The grade 3 and grade 4 results were compared using Pearson's Chi-squared test.

For the addition and subtraction problems, the proportion of correct answers was nearly perfect in both grades. The differences between grade 3 and grade 4 were not significant in either case, χ^2 (1, N = 651) = 2.11, p = .146, φ = –.06; and χ^2 (1, N = 651) = 0.40, p = .528, φ = –.03. This suggests that these operations are already well established in grade 3. In the case of multiplication, the proportion of correct solutions was somewhat lower (≈ 88%). However, the difference between grade 3 and grade 4 was not significant, χ^2 (1, N = 651) = 0.61, p = .434, φ = .03, indicating that multiplication performance approaches the stability of addition and subtraction more gradually. In contrast, division showed marked development: 86.3% of grade 3 students and 96.9% of grade 4 students responded correctly. The difference was significant, χ^2 (1, N = 651) = 22.59, p < .001, φ = .19, indicating a small-to-medium effect size (Cohen, 2013). This finding suggests that mastery of division typically becomes secure by grade 4, although the effect size indicates that the grade level explains only part of the variance.

Next, a descriptive statistical analysis based on the scores obtained for the survey (**Table 5**).

The maximum score was 12 (1 or 0 points for each WP). The mean of the overall scores is 8.38 (standard deviation = 2.68). The median is close to this, at 9, while the first quartile is 7 and the third quartile is 10. It means that, on average, the participants answered 8-9 WPs correctly. Looking at the performance of third and fourth graders separately, we see that the average number of correct answers was 7-8 and 9-10, respectively. Using the Mann-Whitney U test, we found a significant difference between them; fourth graders performed better (U = 3450, Z = 7.584, p=.0014, r = .30, 95% confidence interval [CI] [.23, .37]). The medium effect size (r = .30, 95% CI [.23, .37]) indicates that this developmental gain is not only statistically significant but also reflects a tangible performance improvement.

The distribution of students' overall scores is shown in **Figure 1**. The most frequent scores are 11 for fourth graders and 8 for third graders.

Table 6 gives an overview of percentages of correct answers for each problem. The grade 3 and grade 4 results are also compared. The comparison shows that fourth graders scored significantly better than third

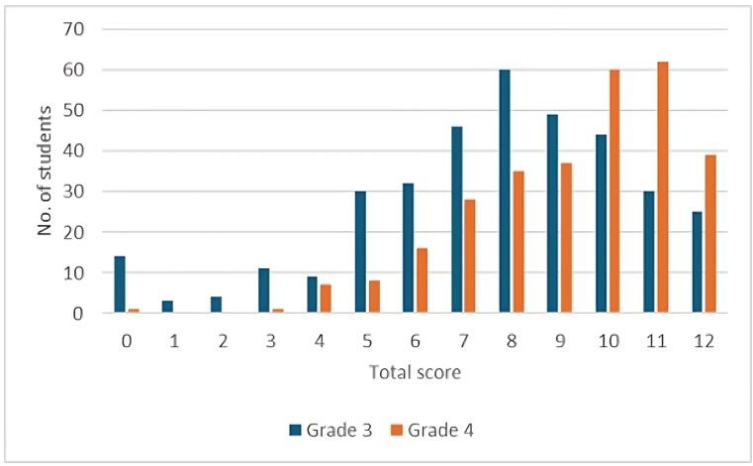

Figure 1. Students' overall performance (Source: Authors' own elaboration)

Table 6. Percentage of correct answers per task

	Total		Grade 3		Grade 4		Comparison	
	n	%	n	%	n	%	φ	p
P1	606	93.1	319	89.4	287	97.6	.16	< .001
P2	398	61.1	193	54.1	205	69.7	.16	< .001
P3	564	86.6	291	81.5	273	92.9	.17	< .001
P4	613	94.2	324	90.8	289	98.3	.16	< .001
P5	503	77.3	256	71.7	247	84.0	.15	< .001
P6	526	80.8	264	73.9	262	89.1	.19	< .001
P7	294	45.2	129	36.1	165	56.1	.20	< .001
P8	467	71.7	235	65.8	232	78.9	.14	< .001
P9	223	34.3	101	28.3	122	41.5	.14	< .001
P10	535	82.2	272	76.2	263	89.5	.17	< .001
P11	261	40.1	113	31.7	148	50.3	.19	< .001
P12	471	72.4	238	66.7	233	79.3	.14	< .001

Note. N (Total) = 651; N (Grade 3) = 357; N (Grade 4) = 294. Person's Chi-squared test with φ as effect size value

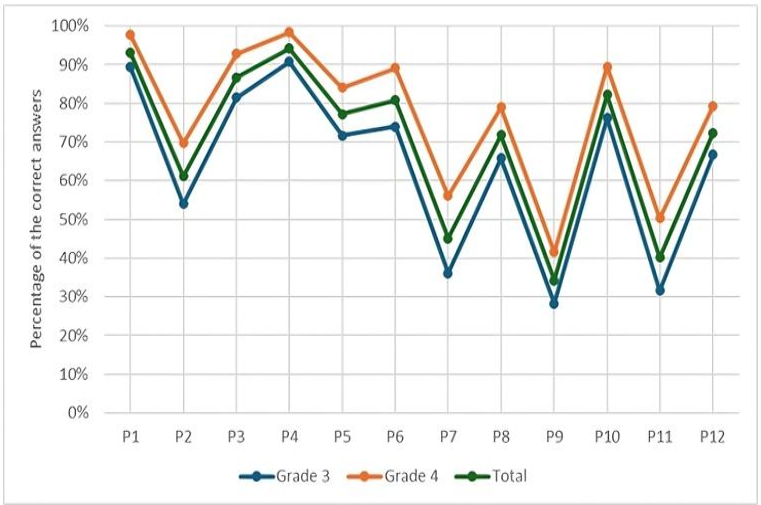

Figure 2. Percentage of correct answers per task (Source: Authors' own elaboration)

graders on all WPs, although the effect sizes range between .14 and .20, which is considered rather small.

Figure 2 also represents students' performance by tasks, where the percentage of the correct answers is indicated on the vertical axis.

It can be seen that the same WPs proved to be easy or difficult for both grades. The biggest difference is in the answers to P7. Based on the data above, the WPs can be ranked in order of difficulty as follows: P4 < P1 < P3 < P10 < P6 < P5 < P12 < P8 < P2 < P7 < P11 < P9. Students in the survey scored the lowest in P7, P9, and P11, with percentages of correct answers below 50%.

Table 7. The result of the problem groups comparisons

	Problem groups	Mean	Standard deviation	Z	p	r	Difference
(1)	Additive (P1, P2, P3, P5, P8)	.78	.25	17.07	< .001	.67	Significant
	Multiplicative (P7, P9, P10, P11, P12)	.55	.28				
(2)	Lexical consistency (P1, P3, P5, P10, P12)	.82	.24	18.57	< .001	.73	Significant
	Lexical inconsistency (P2, P7, P8, P9, P11)	.50	.31				
(3)	Unknown CR (P5, P8, P11, P12)	.65	.30	16.03	< .001	.63	Significant
	Unknown CS (P1, P3, P10)	.73	.24				
	Unknown CR (P5, P8, P11, P12)	.65	.30	12.38	< .001	.49	Significant
	Unknown RS (P2, P7, P9)	.47	.35				
	Unknown CS (P1, P3, P10)	.73	.24	18.77	< .001	.74	Significant
	Unknown RS (P2, P7, P9)	.47	.35				
(4)	Order-consistent (P1, P2, P8, P9, P10, P11)	.64	.25	7.28	< .001	.29	Significant
	Order-inconsistent (P3, P5, P7, P12)	.70	.27				

Note. N = 651 & All analyses were carried out using the Wilcoxon signed-rank test, except for (3), which was carried out using Friedman ANOVA, χ^2 (2, 651) = 561.18, p < .001, with the Wilcoxon signed-rank test as post hoc analysis

Comparison of Problem Groups According to Four Factors

As our research mainly focuses on compare WPs, the two non-compare WPs (P4 and P6) are excluded from further analysis. The ten compare WPs are examined along four factors:

1. operation,
2. lexical consistency,
3. role of unknown, and
4. order-consistency.

According to these factors, the problems are divided into two groups, and the performance achieved in each problem group is compared. **Table 7** presents the results.

As a result of the Wilcoxon tests, we identified significant differences for all four factors.

1. Respondents were significantly better at solving additives than multiplicative WPs (Z = 17.00, p < .001, r = .67). This large effect size confirms that multiplicative thinking causes significant difficulties regardless of the other factors.
2. The lexically consistent WPs were significantly easier than inconsistent ones (Z = 18.50, p < .001, r = .73). The very strong effect indicates that the wording heavily guides students' interpretation, and when these cues conflict with the correct operation, performance drops markedly.
3. When the tasks were grouped according to the role of the unknown, the order of difficulty from easiest to hardest was as follows: CS < CR < RS, with all pairwise differences significant (Z = 16.00, p < .001, r = .63; Z = 12.30, p < .001, r = .49; Z = 18.70, p < .001, r = .74). This finding replicates earlier results (e.g., Carpenter et al., 1981; Daroczy et al., 2015) and confirms that RS problems pose the greatest challenge, presumably because they require reorganization of the relational structure before an operation can be chosen.
4. Order-consistent WPs were more complicated than order-inconsistent ones (Z = 7.28, p < .001, r = .29). Although this effect was small, it suggests that order-consistency does not always facilitate processing; in some cases, order inconsistency may reduce, for example, the need to override misleading linguistic cues.

The analysis of the problem groups showed that there are significant differences in all four factors. Accordingly, we can identify which factor variable results in a harder-to-solve compare problem. These are the multiplicative, lexically inconsistent, RS-type and order-consistent problems.

Since the additive/multiplicative factor showed a fairly significant difference, the above comparisons were also performed for the other three factors, separately for additive and multiplicative WPs. The analyses revealed that within both groups the same hierarchy of difficulty emerged: lexically consistent problems were easier than inconsistent ones; tasks with the CS as the unknown were easier than those with CRs, which in turn were easier than those with the RS; and order-inconsistent problems were solved more successfully than order-consistent ones. Thus, the factors influencing problem difficulty appear to operate similarly across additive and multiplicative domains. The detailed statistical results supporting these findings are provided in **Appendix A** (**Table A1**, **Table A2**, **Table A3**, and **Table A4**).

Problem Pairs Which Differ in Only One Factor

After comparing the problem groups set on different factors, pairwise analyses are also carried out to further explore the difficulties arising from the fine structure of the compare WPs.

The proportion of correct answers to problem pairs, which differ in only one factor (**Table 3**) was investigated on the whole sample.

Table 8. Additive vs. multiplicative problem pairs

	Problem pairs	Mean	p	Difference
(1)	P2–Additive	.61	< .001	Significant
	P9–Multiplicative	.34		
(2)	P1–Additive	.93	< .001	Significant
	P10–Multiplicative	.82		
(3)	P8–Additive	.72	< .001	Significant
	P11–Multiplicative	.40		
(4)	P5–Additive	.77	0.02	Non-significant
	P12–Multiplicative	.72		

Table 9. Problem pairs where the unknown is the relation or one of the sets

	Problem pairs	Mean	p	Difference
(1)	P2–Unknown RS	.61	< .001	Significant
	P8–Unknown CR	.72		
(2)	P9–Unknown RS	.34	.014	Non-significant
	P11–Unknown CR	.40		
(3)	P3–Unknown CS	.87	< .001	Significant
	P5–Unknown CR	.77		

What Operation is Required?

Our survey includes four pairs of compare WPs that differ only in the additive/multiplicative factor. As a result of the comparison using McNemar tests, we found that students performed better at solving the additive than the corresponding multiplicative compare WP in all four cases (**Table 8**). A significant difference was identified in three of the four problem pairs.

(1) P2 and P9 are lexically inconsistent problems; the unknown is the RS, and they are order-consistent. However, P2 is an additive while P9 is a multiplicative problem. The ratio of correct answers for the additive problem (P2) was significantly higher than the multiplicative one (P9), χ^2 (1, N = 651) = 106.98, p < .001, OR = 4.24, 95% CI [3.15, 5.70]. This large difference indicates that multiplicative reasoning is substantially more demanding than additive reasoning under these conditions.

(2) P1 and P10 are lexically consistent problems; the unknown is the CS, and they are order-consistent. However, P1 is an additive problem, whereas P10 is a multiplicative one. The additive version (P1) proved significantly easier than the multiplicative one (P10), χ^2 (1, N = 651) = 42.61, p < .001, OR = 4.23, 95% CI [2.66, 6.73]. This finding shows that even in lexically consistent and order-consistent cases, multiplicative problems still pose the greater challenge.

(3) P8 and P11 are lexically inconsistent problems, the unknown is the CR, and they are order-consistent. The difference was even more pronounced: the additive P8 was solved correctly far more often than the multiplicative P11, P11; χ^2 (1, N = 651) = 156.81, p < .001, OR = 7.65, 95% CI [5.26, 11.12]. This pair revealed the largest gap (OR = 7.65), highlighting that the multiplicative problem is especially difficult, far more so than their additive counterparts, when the CR is unknown and the wording is inconsistent.

(4) P5 and P12 are lexically consistent problems, the unknown is the CR, and they are order-inconsistent Similar to the first three problem pairs, the additive version (P5) was somewhat easier, but the difference did not reach the adjusted significance threshold, χ^2 (1, N = 651) = 5.22, p = .022, OR = 1.42, 95% CI [1.06, 1.91]. This suggests that in order-inconsistent problems, additive and multiplicative reasoning may present a more comparable level of difficulty.

Taken together, these comparisons demonstrate that across a variety of structural conditions, additive compare WPs are consistently easier for students than multiplicative ones, with the performance gap being especially pronounced when lexical inconsistency, order-consistency and CR-unknown are combined. In contrast, the difference was the smallest in the case of the lexically consistent, order-inconsistent, and CR-unknown pair.

What is the Unknown?

Three problem pairs are examined in more detail. In each pair, the problems differ in the role of the unknown. **Table 9** contains the results of the McNemar test.

Across the examined problem pairs, a consistent pattern emerged: problems with an unknown RS were systematically more difficult than those with an unknown CR or CS, supporting the hypothesized ranking of difficulty (CS < CR < RS). Our findings confirm that the role of the unknown exerts a robust effect on problem difficulty.

(1) P2 and P8 are additive WPs, they are lexically inconsistent and order-inconsistent. However, in P2, the unknown is the RS, while P8 is the CR. Students performed significantly worse on P2 than on P8, χ^2 (1, N = 651) = 20.1, p < .001, OR = 0.49, 95% CI [0.37, 0.66], g = .34. The large effect size underscores that identifying the RS is more difficult.

(2) P9 and P11 are multiplicative WPs, they are lexically inconsistent and order-consistent. As in the previous problem pair, the RS in P9 is unknown, while the CR is in P11. The comparison's result is similar: P11 (CR), proved easier than P9 (RS), although the effect was smaller: χ^2 (1, N = 651) = 5.9, p = .015, OR = 0.71, 95% CI [0.55, 0.93], g = .17. This medium-sized effect may indicate that, in addition to the difficulty arising from RS, the multiplicative

Table 10. Order-consistent vs. order-inconsistent problem pairs

	Problem pairs	Mean	p	Difference
(1)	P1–Consistent	.93	< .001	Significant
	P3–Inconsistent	.87		
(2)	P7–Inconsistent	.45	< .001	Significant
	P9–Consistent	.34		

operation itself and also the lexical inconsistency pose a considerable challenge.

(3) P3 and P5 are additive WPs, they are lexically consistent, but order-inconsistent. In P3, the unknown is the CS, while in P5, it is the CR. Students performed significantly better on P3 (CS) than P5 (CR), χ^2 (1, N = 651) = 34.7, p < .001, OR = 2.39, 95% CI [1.68, 3.39], g = .41. This large effect size confirms that the CS-type problem is the easiest one.

What is the Order of Numbers in the Text and the Operation?

Two problem pairs differ in the order of the numbers. **Table 10** contains the results of the McNemar tests.

(1) P1 and P3 are additive WPs; they are lexically consistent, and the unknown is the CS. Students performed significantly better on the order-inconsistent problem (P3) than on the order-consistent problem (P1), χ^2 (1, N = 651) = 20.01, p < .001, OR = 0.33, 95% CI [0.20, 0.55], φ = .50. This result suggests that in additive problems with lexical consistency, order-consistency may facilitate correct reasoning, contrary to expectation based on the result of group comparison.

(2) P7 and P9 are additive WPs; they are lexically inconsistent, and the unknown is the RS. P9 is order-consistent, while P7 is order-inconsistent. The performance was significantly higher on the order-inconsistent problem (P7), χ^2 (1, N = 651) = 19.84, *p* < .001, OR = 0.55, 95% CI [0.43, 0.72], φ = .29. This finding indicates that when lexical inconsistency is combined with a reference-set unknown, students may benefit from order inconsistency.

As a result of the tests, we identified significant differences between the mean scores of the problems in both pairs. However, it is interesting that the significantly better results of students in the first and second pair of tasks were different: in the first pair, they managed better to complete the problem characterized by the same orders, while in the second pair, the opposite was true. The latter aligns with our findings regarding the problem group based on order-consistency.

These results reveal that order inconsistency did not increase difficulty as hypothesized. The success depends much more on other factors than on order-consistency, which points to a more complex interaction between order and structural factors than predicted.

DISCUSSION

12 one-operator WPs, each characterized by five factors, were used for our survey (semantic category, operation, lexical consistency, role of unknown, and order-consistency). Compare WPs were divided into groups and then pairs based on these factors. The average performance of 3rd and 4th grade students and differences between the mean scores of each group and pair of problems were analyzed. The findings reveal both expected and unexpected patterns, with significant implications for both theory and practice.

Since the answer to a WP was limited to identifying the arithmetic operation, it was assessed beforehand whether the students could perform the four operations required. These results confirm that students were generally able to carry out the operations involved. At the same time, the high level of computational fluency supports the assumption that, while this check primarily demonstrates proficiency, it does not contradict–and may even reinforce–the possibility that students were also aware of the outcomes of the operations when selecting the one they considered appropriate for the given problem.

The average performance of the students on the 12-point survey was 8.38, corresponding to more than 75% correct responses. While this does not in itself rule out the possibility of guessing, it suggests that students generally made considered selections, going beyond random choice (chance level = 25%). Fourth graders performed slightly better (mean = 9.27) than third graders (mean = 7.66), which is consistent with expectations. A year of additional practice appears to have a positive impact on performance, even though the problems were built on content introduced in the second grade.

The number of correct answers measured the difficulty of the WPs. The easiest was the one with the highest number of correct answers, i.e., the one with the highest mean score. When looking at the mean scores per WP, it was found that the difficulty of the WP did not depend on the grade level at which the students were learning. The same problems were challenging for third and fourth-graders. This suggests that the changes in the mean scores depend on the characteristics of the WPs; thus, these values indicate the WPs' difficulty level. It was also observed that fourth graders performed better in all the problems, which can be explained by the fact that students' results improved without targeted intervention in the research area, thanks to a year of extra schooling.

Based on the above, the WPs were ranked in order of difficulty. The three easiest problems were P4, P1, and

P3, with a success rate above 85%. All of them are additive WPs with lexical consistency. Moreover, the three most difficult problems, which fewer than half of the students could solve, were P7, P11, and P9, namely multiplicative compare problems with lexical inconsistency. Given the known research results (Hegarty et al., 1992; Verschaffel, 1994), it is unsurprising that additive lexically consistent WPs are at the top of the ranking while multiplicative lexically inconsistent WPs are at the bottom.

Nunes et al. (2016) note in their book that compare problems are more challenging than, for example, change problems among additive problems. Our ranking supports this: P4, the single change problem, was the easiest. Among the multiplicative problems, however, findings are not entirely the same. P6, the only basic (i.e., not compare) multiplicative problem, which falls into the equivalent groups' category, is in the first half of the rank. Still, we found one multiplicative WP, which is easier than it. P10 is a simple multiplicative compare problem requiring the calculation of a third of a quantity. Therefore, the difficulty level of multiplicative compare problems should also depend on other factors.

To explore the difficulties inherent in each problem in more depth, problem groups and pairs were constructed and compared according to the different factors.

First, WPs from an additive/multiplicative perspective were investigated and found that additive WPs are significantly more straightforward, which is well-known in the literature (Nunes et al., 2016). This is due to the difficulty of interpreting multiplicative operations, their cognitive complexity, and the one-year delay in learning them. Students' learning experience in the recent sample is one year less (started in the second school year) than that of the additive WPs (started in the first school year). The survey findings indicate that even when students are computationally fluent, multiplication and division continue to pose greater challenges in constructing a mental model of a situation.

The result comparing lexically consistent/inconsistent problems, on the one hand, confirms previous research findings that inconsistent additive WPs are more complicated than consistent ones (Nunes et al., 2016). On the other hand, similar differences were detected in the domain of multiplicative WPs, more narrowly, on multiplicative compare WPs. This aligns with the consistency hypothesis (Lewis & Mayer, 1987), according to which problem solvers perform better when relational terms match the required operation. These data, together with recent evidence from eye-tracking (Bartalis et al., 2023), suggest that lexical cues are highly influential in shaping problem interpretation. The authors of this paper agree with Mayer and Hegarthy's (1996) explanation that, in many cases, the reason for incorrect answers to lexically inconsistent problems may be using a direct translation strategy, whether the problem is additive or multiplicative.

The easiest of the additive compare WPs, according to several studies (Carpenter et al., 1981; Nunes et al., 2016; Riley et al., 1983), are those in which the CS is the unknown, followed by those in which the CR while the most difficult are problems in which the RS is the unknown. To relate these results to the present study, the additive and multiplicative problems were also examined separately from this point of view. The same order of difficulty (CS > CR > RS) was observed for both additive and multiplicative problems, and the difference in the number of correct answers proved significant in all cases (see **Table A2** and **Table A3** in **Appendix A**). This new result confirms that the difficulty of the compare problem is strongly influenced by which of the three quantities is unknown, regardless of the type of operation.

Finally, within the compare WPs, it was investigated whether order-inconsistent problems are more complex than order-consistent ones. Contrary to expectations based on previous studies of additive WPs (Daroczy et al., 2015), order-inconsistent problems were found significantly easier than order-consistent ones. This pattern also emerged when the additive and multiplicative problems were investigated separately (**Table A4** in **Appendix A**). This means that the assumption that the same order results in an easier problem to solve in general was not proven true; a more detailed analysis of the problems in question can investigate why. One possible explanation is that in order-inconsistent problems, students are forced to process the relational structure more carefully, which may lead to fewer errors than when they rely on surface-level order cues. In the next section, this issue is analyzed more deeply.

When the ten compare WPs were divided into two groups based on a particular factor, attention was inevitably focused on that factor, which could obscure the role of the others. To consider the difference between the effects of each factor, the success rates of pairs of problems differing in only one of the four factors were investigated.

Four pairs of tasks were included in the survey, differing only in that one could be solved by an additive operation and the other by a multiplicative operation. The pairwise comparison clearly showed that this difference is very crucial for the difficulty of the compare problems. The multiplicative WP proved more complicated than the additive one in all pairs. This was true when the pair was more challenging regarding the other factors (P2-P9) and when it was easier (P1-P10). This finding is consistent with the different cognitive complexity of additive and multiplicative structures. It should be noted, however, that for the P5-P12 pair, the

difference did not reach significance at the level of □ = .01. These are lexically consistent problems in which the unknown is the CR, and proved relatively easy, with 77% (P5) and 72% (P12) of students responding correctly. The lack of a significant additive–multiplicative difference is likely due to the wording of the tasks: the questions 'how many more?' and 'what fraction of?' provide explicit cues to the required operation, thereby reducing the potential for error. However, it may be necessary to repeat the survey for more groups of students. If the same is found, searching for and investigating additional factors could help uncover the phenomenon's cause. Nevertheless, this is beyond the scope of the current study.

Three other pairs differ only in the type of unknown. The P2-P8 and P9-P11 pairs are lexically inconsistent; the former is additive, and the latter is multiplicative. For both pairs, we found that the WPs in which the RS was unknown were more challenging than those in which the CS. This implies that the unknown type is dominant when the WPs are harder (P9-P11) or easier (P2-P8) in the other factors. For the third pair of problems (P3-P5), P5 was significantly more difficult. In P5, the relation is unknown (CR), while in P3, the CS. They were additive, lexically consistent tasks (i.e., easier regarding these factors), and order-inconsistent. In other words, we confirm the result from the literature for additive tasks (CS > CR > RS) by adding that we obtain a similar result for the multiplicative pair of compare WPs. So, the order of difficulty does not change if the additive compare problems are simpler or more complex in terms of other factors. Note, however, that the difference between the multiplicative pair (P9-P11) was non-significant at the level of □ = .01. Repeating the survey and, if the result is similar, a deeper analysis of the problem pair's text may provide a more convincing answer to the question of whether the observed order of difficulty also exists among tasks that are considered difficult based on all other factors.

The order-inconsistent problem was found to be significantly more difficult for the additive, lexically consistent, compare WPs where the unknown was the CS (P1-P3), suggesting that the order-consistency may indeed influence the difficulty of the tasks, at least for those that appear simpler in terms of other factors. In contrast, P7 and P9 are the most challenging tasks because they are multiplicative, lexically inconsistent WPs where the unknown is the RS. The result shows that P9 is significantly more difficult than P7, even though P9 is order-consistent, while P7 is not. This means that for this pair of problems, the order-consistency as a difficulty factor had the opposite effect than expected. Together with the findings of problem group comparison, this leads to the assumption that this factor is not as decisive as the other three, and additional factors must influence the difficulty. To explain this, let us examine the two pairs more deeply, focusing on the missing terms of the operation in the arithmetic sentences and the order in which the necessary information is detected when reading.

P1. Boti collected 33 chocolate eggs at Easter, and Ákos collected 3 more eggs than Boti. How many chocolate eggs did Ákos collect?

P3. Mom is 3 years younger than dad. How old is mom if dad is now 33?

Reading the text of P1 word by word, we can directly write down the arithmetic sentence needed to solve it: 33+3=□. The unknown occurs only in the question at the end of the text. In contrast, the wording of P3 is more complicated because we must read the whole text to get the first term (33) of the arithmetic sentence: 33-3=□. This confirms that the order-inconsistent P3 is more difficult than the order-consistent P1.

P7. Nóri has three times as many fives (the highest grade in Hungarian schools) as Flóra. How many fives does Flóra have if Nóri has 33?

P9. My brother and I collected seashells during the holiday. My brother collected 33 shells, a third of my number of shells. How many shells did I collect during the holiday?

As in the case of P3, to construct the arithmetic model of P7, we must read and understand the whole text: 33=3□□. Moreover, the unknown (Flóra's fives) is already present in the text before the question. The same is true for P9 concerning the appearance of the unknown (my number of shells), but the arithmetic sentence can be formed by reading the text word by word 33=□□3. Nevertheless, the experience is that P9 was significantly more complex than P7. The challenge of interpreting the keywords could explain this. In P7, this is "three times"; in P9, "a third of". It is known that division is the operation that causes students to have the most problems with interpretation and application, and the keyword in P9 refers to division (Pape, 2003). In addition, the linguistic formulation of P9 is more complex since the quantities involved in the multiplicative comparison are referred to by personal pronouns instead of persons' names.

Taken together, the results identify the main sources of difficulty in compare WPs: multiplicative operations, lexical inconsistency, reference-set unknowns, and, in some cases, order-consistency. From a pedagogical perspective, these findings underscore the importance of explicitly teaching students how to interpret relational language and reason flexibly about the different roles of the unknown. From a STEM perspective, compare problems offer a natural context for fostering quantitative reasoning: they require identifying quantities in everyday or scientific situations, recognizing their relationships, and expressing these relationships mathematically. Awareness of the sources of linguistic and structural difficulties can thus help teachers in STEM education to clearly and easily express

relationships between quantities. This can strengthen students' ability to integrate mathematical modelling with scientific reasoning, which is an important competency in elementary school STEM education.

CONCLUSION

In our research, we investigated the effect of four linguistic and numerical factors on the difficulty level of compare WPs. The sample consisted of 3rd and 4th grade students from different schools in Eastern Hungary, with 651 students taking a written survey. Finding the arithmetic model for the single-operation problems by including the same two numbers in each WP were asked for. The difficulty of the tasks was characterized by the mean scores obtained by the students. We answer the research question, "to what extent do the type of operation, lexical consistency, role of the unknown, and order-consistency affect the success rate of third and fourth graders in finding the arithmetic model of compare WPs?" by comparing the findings with the literature.

Well-established factors such as operation type, lexical consistency and the unknown quantity in the comparison are good predictors of the order of difficulty of WPs. Multiplicative problems were found to be significantly more difficult than additive ones. Lexical consistency is also a crucial factor. Lexically inconsistent problems were more challenging concerning both additive and multiplicative problems. These results are in line with previous research referred to in the literature. It is also known that the difficulty level of compare additive WPs depends on the role of the unknown quantity. We confirmed the known CS > CR > RS order for additive compare problems and even confirmed it for multiplicative problems. However, the fourth factor included in our analyses, the so-called order-consistency, did not appear to be as dominant as the previous ones in terms of difficulty. Contrary to expectations, the order-consistent problems were significantly more difficult than the order-inconsistent ones. This is particularly true for the more challenging problems concerning the other factors. Exploring the interaction between the factors and examining other linguistic factors (sentence structure, grammatical elements) may further explain our ranking.

The findings underscore that solving WPs is not merely a matter of computational fluency but requires the integration of linguistic comprehension, relational reasoning, and arithmetic modelling. For pedagogy, this implies that explicit attention should be given to the linguistic formulation of problems and also supporting the gradual development of multiplicative reasoning. For research, the results highlight the importance of examining interactions between factors, especially the surprising advantages observed for order-inconsistent problems. Overall, compare WPs provide a valuable setting for developing quantitative reasoning and thus contribute to building a foundation for integrated STEM learning in primary education.

Limitation

These findings should be interpreted with caution, considering several design constraints. First, the use of fixed operands (33 and 3) may have limited the generalizability of the results, as different numerical values could yield different patterns of performance. Second, the preliminary computational fluency check, although necessary to establish students' proficiency, may have primed them toward certain outcomes, influencing subsequent choices. Third, the nested data structure was not fully accounted for. While the analyses relied on pairwise non-parametric tests treating responses as independent, this simplification may have led to underestimated variances and thus an inflated risk of type I error. Nevertheless, the relatively large and diverse sample provides some robustness, and the approach has the advantage of keeping the results interpretable for educational practice. Finally, the result obtained may also depend on the sample examined; however, the planned teaching experiment will be implemented on the same or a similar sample, which makes the recent research reasonable. For these reasons, the present conclusions should be regarded as robust associations rather than causal claims, and future studies using more varied designs and multilevel modelling are needed to confirm and extend these results.

Author contributions: EK: conceptualization, methodology, validation, investigation, writing, visualization, funding acquisition; **BV:** formal analysis, data curation, visualization. Both authors agreed with the results and conclusions.

Funding: This study was funded by the Research Program for Public Education Development of the Hungarian Academy of Sciences (KOZOKT2022).

Ethical statement: The authors stated that ethical approval was waived as the survey presented only asked students to solve mathematical problems. The survey was part of a school mathematics class, and students' contributions (i.e., answering math questions) did not differ from usual mathematics class activities. The authors further stated that students' responses were stored and processed anonymously. No personal data or opinions were collected. Written permission was obtained from the principals of the participating schools to conduct the survey.

AI statement: The authors stated that generative AI tools (i.e., ChatGPT) were used solely to improve language clarity and style during manuscript preparation. No content, analyses, or interpretations were generated by AI; these remain entirely the responsibility of the authors. The final version was reviewed and approved in full by the authors.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Bartalis, Á., Péntek, I., & Zsoldos-Marchiș, I. (2023). A pilot study on investigating primary school students' eye movements while solving compare word problems. *Open Education Studies, 5*(1), Article 20220207. https://doi.org/10.1515/edu-2022-0207

Carpenter, T. P., Hiebert, J., & Moser, J. M. (1981). Problem structure and first-grade children's initial solution processes for simple addition and subtraction problems. *Journal for Research in Mathematics Education, 12*(1), Article 27. https://doi.org/10.2307/748656

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. *Cognition and Instruction, 8*(3), 261-289. https://doi.org/10.1207/s1532690xci0803_2

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H.-C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00348

de Corte, E., & Verschaffel, L. (1987). The effect of semantic structure on first graders' strategies for solving addition and subtraction word problems. *Journal for Research in Mathematics Education, 18*(5), 363-381. https://doi.org/10.5951/jresematheduc.18.5.0363

de Corte, E., Verschaffel, L., & Pauwels, A. (1990). Influence of the semantic structure of word problems on second graders' eye movements. *Journal of Educational Psychology, 82*(2), 359-365. https://doi.org/10.1037/0022-0663.82.2.359

de Corte, E., Verschaffel, L., & Van Coillie, V. (1988). Influence of number size, problem structure, and response mode on children's solutions of multiplication word problems. *The Journal of Mathematical Behavior, 7*(3), 197-216.

Erdész, E., & Kovács, V. (1982). *Matematika–Munkalapok 3. osztály* [Mathematics–Worksheets for grade 3]. Tankönyvkiadó Vállalat.

Greer, B. (1987). Understanding of arithmetical operations as models of situations. In J. A. Sloboda, & D. Rogers (Eds.), *Cognitive processes in mathematics* (pp. 60-80). Clarendon Press.

Greer, B. (1992). Multiplication and division as models of situations. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 276-295). Macmillan Publishing Co, Inc.

Greer, B., Verschaffel, L., & De Corte, E. (2002). "The answer is really 4.5": Beliefs about word problems. In G. C. Leder, E. Pehkonen, & G. Törner (Eds.), *Beliefs: A hidden variable in mathematics education?* (pp. 271-292). Springer. https://doi.org/10.1007/0-306-47958-3_16

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology, 84*(1), 76-84. https://doi.org/10.1037/0022-0663.84.1.76

Herendiné Kónya, E. (2013). *A matematika tanítása az alsó tagozaton* [Teaching mathematics in primary grades. Nemzedékek Tudása Tankönyvkiadó.

Hungarian Government. (2020). 5/2020 (I. 31.) Kormányrendelet a nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról szóló 110/2012. (VI. 4.) Korm. Rendelet módosításáról [Government decree no. 5/2020 (I. 31.) on the revision of the Gov. decree no. 110/2012. (VI. 4.) on the publication, introduction and application of the national core curriculum]. *Magyar Közlöny, 17*, 290-447.

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79*(4), 363-371. https://doi.org/10.1037/0022-0663.79.4.363

Mayer, R. E., & Hegarty, M. (1996). The process of understanding mathematical problems. In R. J. Sternberg, & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29-54). Lawrence Erlbaum Associate Inc.

Nunes, T., Dorneles, B. V., Lin, P.-J., & Rathgeb-Schnierer, E. (2016). *Teaching and learning about whole numbers in primary school*. Springer. https://doi.org/10.1007/978-3-319-45113-8_1

Páchová, A., & Vondrová, N. (2021). The effect of semantic cues on the difficulty of word problems and the interplay with other complicating variables. *Research in Mathematics Education, 23*(1), 85-102. https://doi.org/10.1080/14794802.2020.1867229

Pape, S. J. (2003). Compare word problems: Consistency hypothesis revisited. *Contemporary Educational Psychology, 28*(3), 396-421. https://doi.org/10.1016/S0361-476X(02)00046-2

Pongsakdi, N., Kajamies, A., Veermans, K., Lertola, K., Vauras, M., & Lehtinen, E. (2020). What makes mathematical word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills. *ZDM Mathematics Education, 52*, 33-44. https://doi.org/10.1007/s11858-019-01118-9

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability. In H. P. Ginsberg (Ed.), *The development of mathematical thinking* (pp. 153-196). Academic Press.

Selter, C. (2000). Verschaffel, L., Greer, B., and de Corte, E., making sense of word problems. *Educational Studies in Mathematics, 42*(2), 211-213. https://doi.org/10.1023/A:1004190927303

Szomjú, L., & Habók, A. (2015). Matematikai szöveges feladatok és tanulási szokások kapcsolatának vizsgálata [Investigating the relationship between mathematical word problems and learning habits. *Iskolakultúra, 25*(3), 15-31. https://doi.org/10.17543/ISKKULT.2015.3.15

The Educational Authority Hungary. (2020). *Kerettanterv az általános iskola 1-4. evfolyama számára–Matematika 1-4. Évfolyam* [Framework curriculum for grade 1-4–Mathematics. Oktatas, Hungary. https://www.oktatas.hu/kozneveles/kerettantervek/2020_nat/kerettanterv_alt_isk_1_4_evf

Verschaffel, L. (1994). Using retelling data to study elementary school children's representations and solutions of compare problems. *Journal for Research in Mathematics Education, 25*(2), Article 141. https://doi.org/10.2307/749506

APPENDIX A

Table A1. Comparison of additive and multiplicative problems by lexical consistency

Problem groups	Mean	Standard deviation	Z	p	r	Difference
Lexically consistent–Additive (P1, P2, P8)	.86	.25	-12.37	< .001	.48	Significant
Lexically inconsistent–Additive (P3, P5)	.66	.38				
Lexically consistent–Multiplicative (P9, P10, P11)	.77	.34	-17.00	< .001	.67	Significant
Lexically inconsistent–Multiplicative (P7, P12)	.40	.35				

Note. N = 651 & All analyses were carried out using the Wilcoxon signed-rank test

Table A2. Comparison of additive problems by type of unknown

Problem groups	Mean	Standard deviation	Z	p	r	Difference
Unknown CR (P5, P8)	.75	.36	9.98	< .010	.39	Significant
Unknown CS (P1, P3)	.90	.24				
Unknown CR (P5, P8)	.75	.36	6.87	< .010	.27	Significant
Unknown RS (P2)	.61	.49				
Unknown CS (P1, P3)	.90	.24	13.34	< .010	.52	Significant
Unknown RS (P2)	.61	.49				

Note. N = 651 & All analyses were carried out using Friedman ANOVA, χ^2 (5, 651) = 885.47, p < .001, with Wilcoxon signed ranked test as post-hoc analysis

Table A3. Comparison of multiplicative problems by type of unknown

Problem groups	Mean	Standard deviation	Z	p	r	Difference
Unknown CR (P11, P12)	0.56	0.36	13.00	< .010	.51	Significant
Unknown CS (P10)	0.82	0.38				
Unknown CR (P11, P12)	0.56	0.36	9.40	< .010	.37	Significant
Unknown RS (P7, P9)	0.40	0.38				
Unknown CS (P10)	0.82	0.38	16.47	< .010	.65	Significant
Unknown RS (P7, P9)	0.40	0.38				

Note. N = 651 & All analyses were carried out using Friedman ANOVA, χ^2 (5, 651) = 885.47, p < .001, with Wilcoxon signed ranked test as post-hoc analysis

Table A4. Comparisons of additive and multiplicative problems by order-consistency

Problem groups	Mean	Standard deviation	Z	p	r	Difference
Order-consistent–Additive (P1, P2, P8)	.75	.29	5.37	< .010	.22	Significant
Order-inconsistent–Additive (P3, P5)	.82	.30				
Order-consistent–Multiplicative (P9, P10, P11)	.52	.31	4.60	< .010	.18	Significant
Order-inconsistent–Multiplicative (P7, P12)	.59	.37				

Note. N = 651 & All analyses were carried out using the Wilcoxon signed-rank test