ZDM
https://doi.org/10.1007/s11858-019-01118-9

ORIGINAL ARTICLE

What makes mathematical word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills

Nonmanut Pongsakdi[1] · Anu Kajamies[1,2] · Koen Veermans[1] · Kalle Lertola[3] · Marja Vauras[1] · Erno Lehtinen[1]

Accepted: 8 December 2019
© The Author(s) 2019

Abstract
In this study we investigated word-problem (WP) item characteristics, individual differences in text comprehension and arithmetic skills, and their relations to mathematical WP-solving. The participants were 891 fourth-grade students from elementary schools in Finland. Analyses were conducted in two phases. In the first phase, WP characteristics concerning linguistic and numerical factors and their difficulty level were investigated. In contrast to our expectations, the results did not show a clear connection between WP difficulty level and their other characteristics regarding linguistic and numerical factors. In the second phase, text comprehension and arithmetic skills were used to classify participants into four groups: skilful in text comprehension but poor in arithmetic; poor in text comprehension but skilful in arithmetic; very poor in both skills; very skilful in both skills. The results indicated that WP-solving performance on both easy and difficult items was strongly related to text comprehension and arithmetic skills. In easy items, the students who were poor in text comprehension but skilful in arithmetic performed better than those who were skilful in text comprehension but poor in arithmetic. However, there were no differences between these two groups in WP-solving performance on difficult items, showing that more challenging WPs require both skills from students.

Keywords Word problems (mathematics) · Text comprehension · Arithmetic skills · Item response theory · Difficulty level · Elementary school mathematics

1 Introduction

Word problems (WPs) serve many purposes in mathematics education. They bring variation to the practising of basic mathematical operations and prepare students to use mathematical skills in everyday situations outside the classroom. WPs differ from other mathematical tasks that are presented in mathematical notation because the problem is laid out through text that describes a situation and a question(s) that must be answered by performing a mathematical operation(s) derived from the descriptions in the text (Verschaffel et al. 2000). The text comprehension theory of Kintsch and his collaborators (Kintsch 1998; van Dijk and Kintsch 1983) has been widely applied to describe processes required to solve mathematical WPs (Kintsch and Greeno 1985; Reusser et al. 1990). This theory distinguishes between textbases (networks of propositions within the text) and situation models (model of the situation described in the text) as two aspects necessary for an adequate mental representation of text (Kintsch 1998, p. 107).

Solving WPs is complex, as the (complete) process involves a number of phases (Montague et al. 2014; Niss 2015; Verschaffel et al. 2000). Depaepe and colleagues (2015) reviewed descriptions of different WP-solving processes (e.g., Blum and Niss 1991; Burkhardt 1994; Mason 2001; Verschaffel et al. 2000) and concluded that, essentially, the process comprises six phases that are not necessarily performed sequentially: (1) understanding and defining the problem situation and developing a situation model; (2)

✉ Nonmanut Pongsakdi
nopong@utu.fi

1 Department of Teacher Education, Centre for Research on Learning and Instruction, University of Turku, Turku, Finland

2 Turku Institute for Advanced Studies, University of Turku, Turku, Finland

3 Department of Mathematics and Statistics, University of Turku, Turku, Finland

Published online: 17 December 2019

developing a mathematical model based on a proper situation model; (3) working through the mathematical model to acquire mathematical results; (4) interpreting the results with respect to the original problem situation; (5) examining whether the interpreted mathematical result is appropriate and reasonable for its goal; and (6) communicating the acquired solution of the original WP. According to this description, WP-solving requires students not only to apply mathematical concepts and procedures (e.g., arithmetic relations) but also to construct a mental representation (Verschaffel et al. 2015) that demands different levels of text comprehension (van Dijk and Kintsch 1983; Reusser et al. 1990).

In this study, we focused on the more demanding WPs that cannot be solved without going through important phases of the complex problem-solving process summarized by Depaepe and colleagues (2015). Instead of dealing with the difficulty from the experts' predefined point of view, we focused on the difficulty as it appears in students' performance (see actor-oriented theory by Lobato 2012). This study aims to (a) explore WP-item characteristics regarding linguistic and numerical factors and their difficulty levels, using item response theory (IRT) modelling; (b) examine an association among WP-solving, text comprehension and arithmetic skills; and (c) investigate whether students with different levels of text comprehension and arithmetic skills differ in their performance on WPs at different levels of difficulty.

1.1 Linguistic and mathematical factors contributing to WP difficulty

Previous studies have investigated different factors contributing to WP difficulty. For example, in the 1980s, researchers began investigating the difficulties that students encounter when solving various WPs, starting from simple arithmetic WPs (e.g., change, combine, compare: Carpenter 1985; Cummins et al. 1988; De Corte and Verschaffel 1987; Greer 1987; Riley and Greeno 1988) and progressing to more complex WPs requiring non-routine thinking (Lee et al. 2014; Verschaffel and De Corte 1997). Based on a recent literature review by Daroczy and colleagues (2015), the factors influencing WP difficulty could be distilled into three components: linguistic factors, numerical factors, and interaction between linguistic and numerical factors (e.g., reading direction and numerical process, order of number word system).

Prior studies reported that linguistic factors, such as the number of words in the WP statement, influence the difficulty of WPs (Jerman and Rees 1972). However, later research has shown that superficial textual features such as the number of words hardly explains the difficulty of WPs. For example, Lepik (1990) reported different findings after investigating students' performance on WPs. Linguistic factors, including the length of a WP statement, were not significant predictors of the proportion of correctly solved WPs but were strong predictors of the WP-solving time. In addition to this basic quantitative property of WPs, empirical evidence has convincingly shown that the semantic structure of WPs strongly impacts WP difficulty and the strategies that young children apply when solving arithmetic WPs (Cummins et al. 1988; De Corte and Verschaffel 1987; LeBlanc and Weber-Russell 1996; Riley and Greeno 1988). Semantic structure refers to meaningful relations between the known and unknown sets involved in the WP (e.g., whether the text of a simple additive WP involves a combination of two sets, a dynamic change in a start set, or a comparison between the magnitudes of two sets: see Riley and Greeno 1988 for WPs with different semantic structures). There are evidently significant differences in the probability of solution for problems both within a specific semantic type and between these semantic structure types (Cummins et al. 1988; De Corte and Verschaffel 1987). According to the literature, many young children have difficulty at the stage of comprehending sentences (Cummins 1991), while difficulties for older students may be more closely connected to the overall demands of arriving at an integrated representation of the situation described in text (LeBlanc and Weber-Russell 1996). In WPs, an additional linguistic factor is related to the role of a situation model in comprehending the meaning of the problem text. In routine WPs, it is enough to understand the propositions presented in the text, whereas in WPs requiring non-routine thinking, the construction of an adequate situation model is necessary in order to understand the problem (Kintsch 1998; Reusser 1992). Linguistic factors, such as irrelevant information and implicit information, have been found to influence students' comprehension of WP-situations, which is essential to the construction of situation models. According to Englert and colleagues (1987), irrelevant numerical information negatively impacts students' WP-solving, while irrelevant linguistic information did not affect their performance. Concerning implicit information, researchers suggested that many unsuccessful problem solvers often rely on the direct translation strategy (looking for numbers and keywords) and fail to provide correct answers, especially when problems include important implicit information, which they should infer on the basis of the situation described in the text (Hegarty et al. 1995). Another factor influencing WP difficulty, which can be seen as an extended aspect of a situation model, is the necessity of using realistic considerations requiring a non-direct translation of the situational model into a mathematical one. WPs that demand the use of realistic considerations were reported to be very difficult for many students (e.g., Verschaffel and De Corte 1997; Verschaffel et al. 2000).

Several studies have shown that numerical factors, such as number properties, required operations and number of

solving steps, influence difficulty too. For example, Koedinger and Nathan (2004) investigated the effect of decimal numbers on students' WP-solving performance. Their results indicated that whole-number problems are significantly easier than decimal-number problems. Apart from the effect of number type, the type of operation required (e.g., addition and subtraction; multiplication and division) appears to have an impact on children's solution strategies and varies widely in difficulty (De Corte and Verschaffel 1987; De Corte et al. 1988). Various kinds of arithmetic calculation errors can result from the type of operation required (Kingsdorf and Krawec 2014). In addition to the required operations, the number of solving steps was reported to have an impact on WP difficulty. For instance, Quintero (1983) examined students' problem-solving performance on WPs with a ratio and revealed that two-step ratio WPs are more difficult than single-step ones. Problems can also require mathematical thinking that goes beyond the basic arithmetic, such as combinatorial reasoning, which has proved to be difficult for children (English 2005).

1.2 Associations between WP-solving, text comprehension, and arithmetic skills

A substantial number of studies have examined an association between WP-solving performance, arithmetic and text comprehension skills. For example, Fuchs and colleagues (2006, 2018) reported that arithmetic skills are related to WP-solving performance and can be seen as an essential foundation for WP-solving. However, their studies indicated that, although arithmetic skills are a necessary foundation, they are not sufficient for WP-solving given that WPs also require text processing when constructing a mental representation. Furthermore, this is evident in several studies that have found associations between WP-solving performance and text comprehension even after controlling for general cognitive abilities (e.g., working memory) or other factors that may be involved (e.g., technical reading, calculation skill, gender) (Boonen, de Koning et al. 2016; Boonen, van der Schoot et al. 2013; Swanson et al. 1993; Vilenius-Tuohimaa et al. 2008). Although the association between WP-solving performance, text comprehension, and arithmetic skills has received much attention in previous research (e.g., Boonen et al. 2013, 2016; Fuchs et al. 2006, 2018; Swanson et al. 1993; Vilenius-Tuohimaa et al. 2008), these studies typically investigated WP-solving performance on WPs that have simple semantic structures and did not pay any attention to the differences in the difficulty levels of WPs. There are well-established findings in the literature on how different semantic problem types have different difficulty levels (LeBlanc and Weber-Russell 1996) and result in different errors (Carpenter 1985; Cummins et al. 1988; Greer 1987; Riley and Greeno 1988). However, there is still a lack of studies on more demanding WPs that require non-routine thinking (e.g., including more complex structures, involving different factors contributing to their difficulty).

Various students may experience WP difficulty differently. Moreover, the effects of students' skills and WP characteristics may be interrelated. For example, linguistically rather weak students (poor in text comprehension) may face challenges with linguistically complex WPs (e.g., including semantically complex features, long WP statements) and arithmetically rather weak students with arithmetically complex WPs (Daroczy et al. 2015). This assumption seems reasonable since empirical evidence has clearly shown associations between WP-solving performance, text comprehension, and arithmetic skills (Boonen et al. 2013, 2016; Fuchs et al. 2006, 2018; Swanson et al. 1993; Vilenius-Tuohimaa et al. 2008). However, it raises important questions as to whether it is possible to identify which WPs are linguistically or arithmetically complex and whether both features can explain the difficulty of more demanding WPs.

1.3 Aims

Our aim in this study is to deepen our understanding of the associations of linguistic and mathematical WP characteristics and WP-solving skills. Previous investigations about these associations were mostly conducted with simple arithmetic WPs (e.g., Boonen et al. 2013; Fuchs et al. 2006, 2018; Swanson et al. 1993; Vilenius-Tuohimaa et al. 2008). In this study, the investigation was carried out with more demanding WPs and the item-level difficulty of WPs was scrutinized. This study had a focus on a joint investigation of linguistic and numerical WP characteristics and took into account students' skills in text comprehension and arithmetic, while *previous studies often focused on one or the other aspect or skill* (Daroczy et al. 2015). In identifying the difficulty level of WPs, previous research typically relied on classical test theory (CTT) in which the proportion of individuals answering the item correctly is used as the index for the item difficulty (Finch and French 2015; Parkash and Kumar 2016; Stage 2003). However, item difficulty index derived from CTT is often criticized because it is dependent on the sample (Chalmers 2012; Stage 2003). A widely recommended alternative to CTT is the item response theory (IRT) modelling (De Ayala 2009; Finch and French 2015; Reckase 2009), in which the difficulty level estimated with IRT refers to a probability of a correct response at a given level of participant ability (Finch and French 2015). With IRT, it is possible to obtain item characteristics (e.g., item difficulty level) that are not dependent on the examinee group (Parkash and Kumar 2016; Stage 2003).

The present study has the aim of answering the following research questions:

1. Are there linguistic or mathematical features that explain the level of difficulty of various WPs?
2. Are students' text comprehension, arithmetic, and WP-solving skills correlated with each other?
3. How do different patterns of students' text comprehension and arithmetic skills predict their performance in WPs of different levels of difficulty?

2 Method

2.1 Participants and overall design

Participants comprised 891 fourth-grade students, including 446 boys and 445 girls, from different elementary schools situated in cities, small towns, and rural communities in southern Finland. All of them had Finnish as their mother tongue. All participants were asked to complete text comprehension, arithmetic, and WP-solving tests in a classroom situation as a part of the Quest for Meaning project. The data were partly used in a previous study (Kajamies et al. 2010). The University of Turku's ethical guidelines were followed. Permissions were obtained from both the schools and the students' guardians.

2.2 Measures

2.2.1 Mathematical word problems

Students' WP-solving performance was measured with a WP test containing 15 items (Kajamies et al. 2003; see Appendix 1). These WPs were created in such a way that they could not be solved using straightforward strategies (e.g., by requiring students to develop a proper situation model, avoiding keywords in the WPs, and giving numerical information in a written format). Two WPs (WP6 and WP13) were constructed based on original items used in earlier studies (Verschaffel et al. 2000). The students had no time limit to complete the WP test. All WPs were assessed by giving one point for each correct answer and zero points for an incorrect answer or no response. Cronbach's alpha for the whole test was 0.76. The number of words, irrelevant information, implicit information, the use of realistic considerations, the required problem-solving steps and arithmetic operations, and the use of decimal numbers were all noted in order to investigate WP characteristics that may influence WP difficulty level (see Table 2).

2.2.2 Text comprehension

Text comprehension skills were assessed with the Finnish Standardized Reading Test (Lindeman 1998). The students were given 48 multiple choice questions about the four different texts they had to read. One point was awarded for each correct answer, making a maximum score of 48 for text comprehension. The Kuder–Richardson coefficient of internal consistency (CR20) was 0.87 (Lindeman 1998). Text comprehension was seen as an important measure of the linguistic skills of 4th graders.

2.2.3 Arithmetic skills

Arithmetic skills were measured with a time-limited (10 min) RMAT test (Räsänen 2004). The test begins with simple computations and ends with algebraic tasks. According to Räsänen (1993), the RMAT is comparable to the WRAT-R (Jastak and Wilkinson 1984). Both of them contain similar instructions, but the RMAT closely follows the Finnish mathematics curriculum (e.g., the role of fractions is not emphasized) and includes more computational tasks. Therefore, it can assess more basic arithmetic than the WRAT-R (correlations were 0.547–0.659, $n = 2673$, Räsänen 2004). The total number of correct solutions in the RMAT is here used as an indication of the students' arithmetic skills. The maximum score was 56. The alpha coefficient was 0.92–0.95 (Räsänen 2004).

2.3 Analysis

Analyses used in the present study are separated into two phases. The first phase investigated item characteristics and employed item response theory (IRT) modelling to classify WPs based on their difficulty level. The second phase used k-means clustering to categorize students into groups based on their text comprehension and arithmetic skills. In addition, the one-way analysis of variance (ANOVA) was used to determine whether students with different text comprehension and arithmetic skills differ in their performance in mathematical WP-solving.

2.3.1 Item response theory (IRT)

IRT is widely employed in assessment and evaluation research in the fields of education and psychology. IRT is an approach of testing, which is based on the relationship between participants' performance on a particular test item and the level of his or her performance in general in all items measuring the skill in question. In technical terms, IRT attempts to model individual response patterns by determining how underlying latent trait levels (i.e., ability) interact with the item's characteristics (e.g., item difficulty, discrimination ability) to form an expected probability of the response pattern (Chalmers 2012; Embretson and Reise 2000). In this study, IRT analyses were conducted using the R 3.2.3 with ltm (latent trait models) package,

which was developed to analyze multivariate dichotomous data using latent variable models (Rizopoulos 2006).

One type of IRT models called 2PL model (two-parameter logistic model) was applied to investigate the difficulty level of WPs. It expresses the relationship between individuals' level of the latent trait (his or her WP-solving ability) and the probability of endorsing a given item (answering the WP correctly) in the form of a logistic model (Finch and French 2015). Relative fit indices (AIC, BIC, Item fit) were examined to see whether the model fits the individual items well.

2PL model:

$$P\left(x_j = 1 \middle| \theta, b_j\right) = \frac{e^{a_j(\theta,-b_j)}}{1 + e^{a_j(\theta,-b_j)}}$$

θ = students' ability; a_j = discrimination value of item j; b_j = difficulty level of item j.

2.3.2 Unidimensionality test

For selecting a suitable IRT model, the dimensionality of a set of test items has to be tested. A primary assumption underlying the 2PL model is that there is only one latent trait being measured by the set of items (unidimensionality). There are many ways to test the unidimensionality assumption (see Finch and French 2015; Verhelst 2002). One approach is to use the bootstrap modified parallel analysis test (BMPAT, Finch and Monahan 2008), which was developed based on Horn's (1965) parallel analysis method for indicating the number of factors. The BMPAT works by checking the second eigenvalue of the observed data to see whether it is larger than the second eigenvalue of the data under the assumed IRT model. If the BMPAT test results are statistically significant for the second eigenvalue ($p < 0.05$), it means that the data are not unidimensional (Finch and French 2015).

2.3.3 2PL model and unidimensionality test

Table 1 shows the fit indices for the 2PL model. The results indicate that the 2PL fits all items well, and based on the BMPAT result, the observed data are unidimensional ($p > 0.05$), and the results support the primary assumption underlying the 2PL model.

Table 1 Fit indices and unidimensionality tests for 2PL model

Model	Item fit	AIC	BIC	BMPAT (*p*)
2PL	All items	14,664.36	14,808.13	> 0.05

3 Results

3.1 Difficulty level and WP characteristics concerning linguistic and numerical factors

Item difficulty values estimated by the IRT-analysis and linguistic and numerical factors of WPs were examined, and the results are presented in Table 2 and Fig. 1. According to Finch and French (2015), the item difficulty estimates are centred at 0. Therefore, the negative values indicate relatively easy items, while the positive values represent somewhat difficult items. The order of items was arranged based on their difficulty level, from the easiest item (WP1) to the most difficult one (WP13).

Overall, the results showed that the association between WP characteristics concerning linguistic and numerical factors and their difficulty level is not simple and straightforward. There was no significant correlation between the number of words and the difficulty value ($r(13) = 0.21$, $p = 0.490$). Within these WPs, the need for realistic considerations did not explain the difficulty, because the two WPs requiring realistic consideration (WP6 and WP13) were located at the different ends of the difficulty dimension. Also the existence of irrelevant linguistic or mathematical information was distributed equally with the difficulty dimension of the WPs and did not distinguish between easy and difficult WPs. However, implicit information seemed to explain WP difficulty: the eight WPs with the lowest difficulty value had no implicit information, whereas four of the five WPs with highest difficulty values (WP12, WP5, WP4, and WP10) had implicit information.

Neither the solving steps ($r(13) = 0.22$, $p = 0.468$) nor the number of possible operations ($r(13) = 0.42$, $p = 0.157$) correlated significantly with the difficulty value, but both of the WPs including decimal numbers appeared to be relatively difficult.

The success rates of WP9 (6.6%) and WP13 (7%) were extremely low, and the curves (see Fig. 1) suggest that the two items were very difficult. Therefore, within the model, these items' difficulty could not be properly estimated (extremely high standard errors). None of the aspects used in the analysis (Table 2) explained the extreme difficulty of WP9. One explanation might be that it was the only WP that required combinatorial reasoning. WP13 required deep understanding of the real-life situation described in the problem. Because only a few students could solve these two extremely difficult WPs, they were excluded from further analyses.

For further analyses, WPs were classified as easy (WP1, WP6, WP3, WP14, and WP15) and difficult items (WP2, WP8, WP7, WP12, WP5, WP4, WP11, and WP10) based on their difficulty values in the IRT analyses. Cronbach's alphas for each subgroup were 0.61 and 0.64, respectively.

Table 2 Item difficulty (SE) and other characteristics

Item	Group	Results from IRT	Linguistic Factor	Situation model			Numerical Factor		
		Difficulty value (β)	No. of words (in Finnish)	Irrelevant linguistic and/or numerical info	Implicit info	Realistic considerations	Solving step(s) required	(Possible) Operation(s) required	Decimal numbers included
WP1	Easy	−0.94 (0.09)	23	Linguistic and numerical	No	No	1	×	No
WP6		−0.30 (0.10)	15	No	No	Yes	1	÷	No
WP3		−0.20 (0.09)	32	Linguistic	No	No	1	÷	No
WP14		0.03 (0.06)	54	Linguistic	No	No	3	×, +	No
WP15		0.03 (0.06)	60	Linguistic	No	No	3	×, +	No
WP2	Difficult	0.25 (0.07)	47	Linguistic and numerical	No	No	2	+, −	No
WP8		0.25 (0.09)	28	Linguistic	No	No	3	+, −, x	No
WP7		0.37 (0.08)	37	Linguistic and numerical	No	No	1	+	Yes
WP12		0.55 (0.10)	45	Linguistic and numerical	Yes	No	3	+, −, ÷	No
WP5		0.92 (0.10)	43	Linguistic	Yes	No	2	+, −, x	No
WP4		0.93 (0.10)	26	No	Yes	No	1	÷	No
WP11		1.07 (0.11)	37	Linguistic	No	No	1	×, +	Yes
WP10		1.324 (0.13)	36	Linguistic	Yes	No	3	×, +	No
WP9	Extremely difficult	N/A[a]	34	No	No	No	1	×	No
WP13		N/A[a]	20	No	No	Yes	2	×, ÷	No

[a]Extremely low success rate and cannot be properly estimated

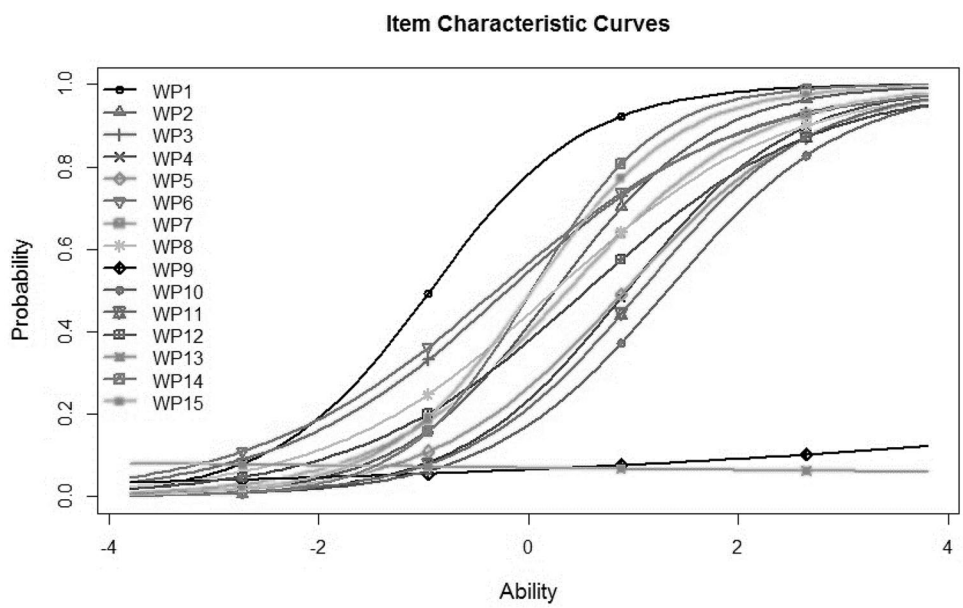

Fig. 1 Item characteristic curves (ICC) of 2PL model

3.2 Associations between text comprehension, arithmetic, and WP-solving skills

To investigate the interrelation among text comprehension, arithmetic, and WP-solving skills on easy and difficult items, Pearson correlations were calculated. The students who did not complete all the three tests were excluded from the analyses (N = 55). The correlation matrix is shown in Table 3, and it revealed a significant correlation between text comprehension skills with both easy ($r(836) = 0.41$, $p < 0.01$)

and difficult items (r(836) = 0.43, p < 0.01). Arithmetic skills also showed a significant correlation with both easy (r(836) = 0.52, p < 0.01) and difficult items (r(836) = 0.53, p < 0.01).

3.3 Individual differences and how those differences relate to WP-solving performance

To investigate individual differences in text comprehension and arithmetic skills and how those differences relate to WP-solving performance, first, students were categorised based on their text comprehension and arithmetic skills. K-means clustering was conducted on the z-scores of the variables. As a result, students were classified into four different groups: poor in text comprehension and arithmetic skills (N = 154); poor in text comprehension but skilful in arithmetic skills (N = 197); skilful in text comprehension but poor in arithmetic skills (N = 288); skilful in both skills (N = 197) (see Fig. 2). Then, ANOVAs were conducted to compare students' text comprehension and arithmetic skills between groups. The results of the ANOVAs indicate that there was a significant difference between groups in their text comprehension (F(3, 832) = 787.666, p < 0.001) and arithmetic skills (F(3, 832) = 519.959, p < 0.001). Post-hoc tests (Bonferroni) revealed that the differences in text comprehension and arithmetic skills were significant in all group comparisons (all ps < 0.001). Descriptive information concerning text comprehension and the arithmetic skills of students in each group is presented in Table 4.

Later, ANOVAs were conducted to investigate whether students in each group differ in their performance in mathematical WP-solving on easy and difficult items. The results of these ANOVAs show that there was a significant difference between groups in mathematical WP-solving performance on both easy items (F(3, 832) = 102.636, p < 0.001) and difficult items (F(3, 832) = 116.554, p < 0.001).

As shown in Table 4, the results of post hoc comparisons using the Bonferroni test reveal that students who were very poor in both skills had the lowest mathematical WP-solving performance on both easy (M = 0.23, SD = 0.25) and difficult items (M = 0.12, SD = 0.14) (all ps < 0.001), while students who were very skilful in both skills had the highest mathematical WP-solving performance on both easy (M = 0.71, SD = 0.23) and difficult items (M = 0.53, SD = 0.22) (all ps < 0.001). In addition, students who were poor in text comprehension but skilful in arithmetic skills (M = 0.53, SD = 0.26) performed significantly better than those who were skilful in text comprehension but poor in arithmetic skills (M = 0.46, SD = 0.27) on easy WPs (p < 0.05).

Table 3 Pearson's correlation matrix between the main variables

Variable	1	2	3	4	5
1. All WPs[a]	–				
2. Easy items	0.86**	–			
3. Difficult items	0.93**	0.61**	–		
4. Zscore_arithmetic	0.59**	0.52**	0.53**	–	
5. Zscore_text comprehension	0.47**	0.41**	0.43**	0.28**	–

*$p < .05$; **$p < .01$; ***$p < .001$

[a]WP9 and WP13 excluded

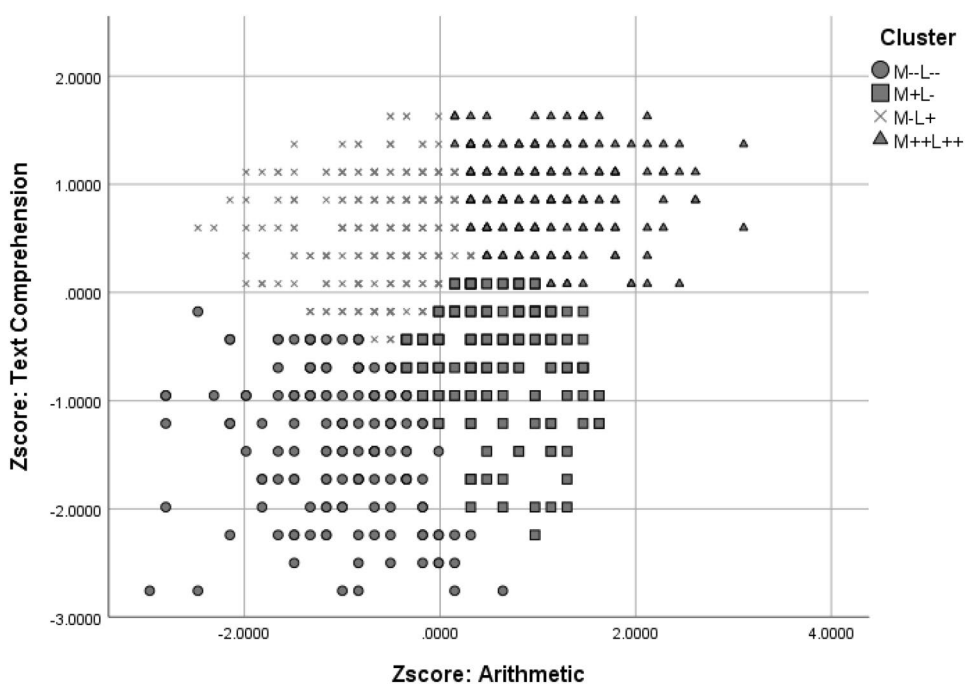

Fig. 2 Four clusters of students based on their text comprehension and arithmetic skills. *M−L−* very poor in both skills, *M+L−* skilful in arithmetic but poor in text comprehension, *M−L+* poor in arithmetic but skilful in text comprehension, *M++L++* very skilful in both skills

Table 4 Descriptive information concerning text comprehension, arithmetic, and WP-solving skills by each student group

Skill	All students N = 836 M (SD)	M−L− N = 154 M (SD)	M + L− N = 197 M (SD)	M− L+ N = 288 M (SD)	M ++L ++ N = 197 M (SD)
Text comprehension	17.69 (3.88)	12.25 (2.56)	15.39 (2.01)	19.77 (1.85)	21.20 (1.59)
Arithmetic	30.09 (6.10)	23.97 (4.18)	33.52 (3.10)	26.53 (3.36)	36.64 (3.90)
WP-solving (all WPs[a])	0.39 (0.23)	0.16 (0.15)	0.42 (0.21)	0.36 (0.20)	0.60 (0.19)
WP-solving (Easy items)	0.49 (0.30)	0.23 (0.25)	0.53 (0.26)	0.46 (0.27)	0.71 (0.23)
WP-solving (Difficult items)	0.33 (0.25)	0.12 (0.14)	0.35 (0.23)	0.30 (0.22)	0.53 (0.22)

[a]WP9 and WP13 excluded

However, there were no differences in these groups' performance on difficult WPs ($p = 0.14$), showing that more challenging WPs require students to also be competent in text comprehension.

4 General discussion

Previous empirical evidence has convincingly shown that WP-solving performance is related to both text comprehension (Boonen et al. 2013, 2016; Swanson et al. 1993; Vilenius-Tuohimaa et al. 2008) and arithmetic skills (Fuchs et al. 2006, 2018). However, these studies mainly examined WP-solving performance on arithmetic WPs with simple semantic structures without paying any attention to the differences in the difficulty level of WPs. This study focused on WP-solving performance when dealing with demanding WPs in which a solution to these WPs requires that students develop a proper situation model and cannot rely solely on superficial coping strategies, such as the keyword approach. The first aim of the study was to investigate WP-item characteristics regarding linguistic and numerical factors and their difficulty level, using item response theory (IRT) modelling. We wanted to find out whether the selected linguistic factors (the length of WP statement), those influencing difficulty in developing situation model (irrelevant information, implicit information, and the use of realistic considerations), and numerical factors (number properties, required operations, and number of solving steps) could explain the difficulty level of these demanding WPs. Overall, the results revealed that superficial linguistic factors did not clearly explain WP difficulty. However, WPs which required inference of implicit information belonged all to the difficult WPs. Both WPs including decimal numbers (WP7 and WP11) were difficult but other numerical factors did not predict the difficulty of WPs. These results are not surprising given that the structure and context of these demanding WPs that require non-routine thinking are fairly different, and there is no common strategy to solve these problems. Individual items seem to have unique factors in their deeper structure that contribute to the item's difficulty level. For example, the main factor that contributes to the difficulty level of the extreme item WP13 seems to be the demand to use real-world knowledge in the modelling process (see Verschaffel and De Corte 1997), while the factor influencing WP difficulty for WP10 could be the difficulty in developing a situation model from the WP statement. One possible explanation for the finding that some of the superficially linguistically similar items appeared to be more difficult is that, in these difficult items, the deep structure is different. For example, the numbers needed in the calculations were not directly given and the students had to infer them from textual expressions.

The second aim of this study was to examine an association between text comprehension, arithmetic, and WP-solving skills on easy and difficult items. In line with prior studies (Boonen et al. 2013, 2016; Fuchs et al. 2006, 2018; Swanson et al. 1993; Vilenius-Tuohimaa et al. 2008), the results showed that text comprehension correlated with WP performance. In this study, we showed that the connection was equally strong with both easy and difficult items. Similar connections also occurred with arithmetic skills; the results indicated that there is an association between arithmetic skills and performance in both easy and difficult items.

The last aim of this study was to investigate individual differences in text comprehension and arithmetic skills and their relationships with WP-solving performance. Students were categorised, based on their text comprehension and arithmetic skills, into four different groups: very poor in both skills; poor in text comprehension but high in arithmetic skills; skilful in text comprehension but poor in arithmetic skills; very skilful in both skills. As expected, the students who were weak in both skills had the lowest mathematical WP-solving performance in both easy and difficult items, while students who were strong in both skills had the highest WP-solving performance in both easy and difficult items. In addition, students who were poor in text comprehension but strong in arithmetic skills performed better than those who were skilful in text comprehension but poor in arithmetic skills in easy WPs. However, there were no differences in the performance of these two groups on difficult WPs. This shows that arithmetic skills are needed in all WPs, but in more challenging WPs, the role of text comprehension skills becomes important as well. These results are

in concordance with previous studies reporting that text comprehension skills have a prominent role in overcoming textual complexities (De Corte et al. 1985; De Corte et al. 1990), for example, when students deal with WPs containing semantically complex features (Boonen et al. 2016).

The main conclusion of the study was that the difficulty of WPs is not based on the surface linguistic features, but there seem to be features in the structure of the WP texts requiring deeper comprehension (Kintsch 1998), which can explain the differences in the levels of difficulty. However, the evidence for this conclusion is not convincing because the WPs used in this study were not systematically planned for the comparison of these features. Future studies are needed with WPs in which surface and deep structure features of the WPs are systematically varied. Further, in investigating individual differences in WP-solving performance, the study focused mainly on text comprehension and arithmetic skills, but other general cognitive abilities (e.g., working memory, motivation) were not included.

Lastly, to examine WP-solving performance and WP difficulty, the study relied merely on WP-test achievement. In future studies, more detailed observations of students' WP-solving processes, for example, through stimulated recall interviews, are needed to understand students' challenges in solving different demanding WPs.

5 Limitations of the study

A major limitation of the study is that in the formulation of WPs, different mathematical and linguistic features were not systematically varied. One consequence of the missing systematic design of the WPs was that it was impossible to make clear theory based sub-categories of the WPs. In addition some widely studied linguistic aspects such as different sematic structures (e.g. LeBlanc and Weber-Russell 1996) were not clearly represented in the WPs. Another limitation is that WP-solving takes much time and it was not possible to collect the data with a larger number of WPs. This aspect also limited the opportunity to find reliable sub-categories representing different mathematical and linguistic aspects of problems.

6 Educational implications

Mathematical WPs can be valuable content for mathematics education, and the use of WPs requiring non-routine thinking and real-world knowledge of the modelling process, instead of mere routine problems, has been recommended by many researchers (CTGV 1992; Mason and Scrivani 2004; Pongsakdi et al. 2016; Verschaffel and De Corte 1997). It is, however, important to be aware of the remarkable differences in the levels of difficulty of the WPs, which are not always self-evident for the teacher or textbook authors. Moving from routine to realistic non-routine tasks also requires novel teaching strategies and teacher scaffolding (Pongsakdi et al. 2016; Pongsakdi et al. 2019). The results of this study show that demanding non-routine WPs require high levels of text comprehension skills. Thus, practising with more demanding WPs is not only beneficial for mathematics learning but can also be an effective way to improve advanced text comprehension skills.

Acknowledgements Open access funding provided by University of Turku (UTU) including Turku University Central Hospital. This research is supported by the Grant 312528 of the Strategic Research Council (Academy of Finland) for the last author in affiliation with University of Turku.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

Appendix 1

See Table 5.

Table 5 Mathematical word problem solving test

	Word problems
WP1	Risto, the neighbour, claims that he can run 53 m in 6 s. You know that you can run 9 m in a second. How many metres can you run in 6 s?
WP2	One summer day, Ville went to the outdoor swimming pool. Ville liked to pool jump more than anything. During the day, he jumped 24 times from the low platform. He had energy to go up to the high 10-m platform only seven times and jump from there. In total, he jumped six times feet first, the rest head first. How many times did he jump head first?
WP3	All of Kasia's six daughters are ill. Kaisa makes daughters healthy soup. She boils the soup—in total, 2 litres and 4 decilitres. How much soup does each girl get when Kaisa divides the soup equally among her daughters?
WP4	Your uncle gives you and two of your cousins 7 marks and 20 pennies for buying candies. You equally divide money before going to the shop. How much money does each one of you get?
WP5	Hanna, Elisa and Niina decided to get money for candy by collecting empty bottles whole morning. They went through all the gardens next to nearby houses. Every girl found 27 bottles from the ground of which nine were broken, in total. They threw the broken bottles away. How many bottles were left for the girls to bring to the shop?
WP6	22 congressmen were taken by taxi to a dinner party arranged by the President. One taxi could hold four passengers. How many taxies were needed?
WP7	Kalle, Toni, Lauri and Antti belong to their class 4 × 100 metres relay team. The teacher was happy about the boys' practice run. Kalle used 13.4 s for his section, Toni 14.1 s, Lauri 14.2 s and Antti 13.1 s. What was the team's time in the relay race?
WP8	During the night, a cat couple went hunting in the garden. They caught, in total, 6 field mice and 16 forest mice. Each one of them ate first 4 delicious forest mice. How many mice did the cats have left for the morning appetite?
WP9	Hannu's home village can be driven by car via four different roads to Tampere. From Tampere can be driven as many as seven different routes to Hannu's new hometown. How many different routes can Hannu travel when he leaves from his home village through Tampere to his new hometown?
WP10	You play basketball with your friends. The opponent's team can make in the first period 24 points. It makes three points less than your own team. Both teams can make the same number of points in both periods. How many points are made in total during the game?
WP11	When you travel to England, you can see street signs where distances are given in miles. One mile is about 1 km 600 m. Imagine that you are near London. On the sign, there is information that London is 11 miles away. How far to London is it in kilometres and metres?
WP12	Jukka, Heikki and Pekka went fishing for a day. They had in total 15 sausages and 32 crackers with them. Jukka and Heikki ate, in total, eight crackers for lunch, but big Pekka needed six crackers. The rest of the crackers, they divided equally for supper together. How many crackers did each boy get in the evening?
WP13	Risto got seven planks, each of which was five metres long. How many 2 m planks could he saw from these?
WP14	The water tank in the garden was completely full when Mum and Kaisa decided to water the rose garden. Mum's watering can could hold 7 litres, and Kaisa's can could hold 4 l. Both filled their cans 3 times from the big water tank and watered the roses. The water tank was finally left with 14 l of water. How much water was in the water tank before Mum and Kaisa started to water the rose garden?
WP15	Mum wanted to get a big basket filled with apples, but she was too tired to collect them alone. Dad and Hannu helped Mum to collect until her basket was full. Dad's basket could hold 6 kilos of apples and Hannu's basket could hold 4 kilos of apples. Both filled their baskets four times and emptied them into Mum's big basket. Mum had earlier collected 13 kilos of apples in her basket. How much were there in Mum's basket in the end?

References

Blum, W., & Niss, M. (1991). Applied mathematical problem solving, modeling, applications, and links to other subjects: State, trends, and issues in mathematics education. *Educational Studies in Mathematics, 22*(1), 37–68. https://doi.org/10.1007/BF00302716.

Boonen, A. J., de Koning, B. B., Jolles, J., & van der Schoot, M. (2016). Word problem solving in contemporary math education: A plea for reading comprehension skills training. *Frontiers in psychology, 7,* 191. https://doi.org/10.3389/fpsyg.2016.00191.

Boonen, A. J., van der Schoot, M., Florytvan, W., De Vries, H., & Jolles, J. (2013). What underlies successful word problem solving? A path analysis in sixth grade students. *Contemporary Educational Psychology, 38*(3), 271–279. https://doi.org/10.1016/j.cedpsych.2013.05.001.

Burkhardt, H. (1994). Mathematical applications in school curriculum. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (pp. 3621–3624). Oxford: Pergamon Press.

Carpenter, T. (1985). Learning to add and subtract: An exercise in problem solving. In E. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 17–40). Hillsdale: Erlbaum.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

CTGV (Cognition and Technology Group at Vanderbilt). (1992). The Jasper series as an example of anchored instruction: Theory, program description, and assessment data. *Educational Psychologist, 27*(3), 291–315. https://doi.org/10.1207/s15326985ep2703_3.

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

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*(4), 405–438. https://doi.org/10.1016/0010-0285(88)90011-4.

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*(348), 1–13. https://doi.org/10.3389/fpsyg.2015.00348.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

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/10.2307/749085.

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77*(4), 460–470. https://doi.org/10.1037/0022-0663.77.4.460.

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. *Journal of Mathematical Behavior, 7*(3), 197–216.

Depaepe, F., De Corte, E., & Verschaffel, L. (2015). Students' non-realistic mathematical modeling as a drawback of teachers' beliefs about and approaches to word problem solving. In B. Pepin & B. Roesken-Winter (Eds.), *From beliefs to dynamic affect systems in mathematics education. Exploring a mosaic of relationships and interactions* (pp. 137–159). Cham: Springer.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Erlbaum.

Englert, C. S., Culatta, B. E., & Horn, D. G. (1987). Influence of irrelevant information in addition word problems on problem solving. *Learning Disability Quarterly, 10*(1), 29–36. https://doi.org/10.2307/1510752.

English, L. D. (2005). Combinatorics and the development of children's combinatorial reasoning. In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning*. New York: Springer.

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York: Routledge.

Finch, H., & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education, 21*(2), 119–140. https://doi.org/10.1080/08957340801926102.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology, 98*(1), 29–43. https://doi.org/10.1037/0022-0663.98.1.29.

Fuchs, L. S., Gilbert, J. K., Fuchs, D., Seethaler, P. M., & Martin, B. N. (2018). Text comprehension and oral language as predictors of word-problem solving: Insights into word-problem solving as a form of text comprehension. *Scientific Studies of Reading, 22*(2), 152–166. https://doi.org/10.1080/10888438.2017.1398259.

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). Oxford: Clarendon Press.

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology, 87*(1), 18–32. https://doi.org/10.1037/0022-0663.87.1.18.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447.

Jastak, S., & Wilkinson, G. (1984). *Wide range achievement test-revised (WRAT-R)*. Wilmington: Jastak Associates.

Jerman, M., & Rees, R. (1972). Predicting the relative difficulty of verbal arithmetic problems. *Educational Studies in Mathematics, 4*(3), 306–323.

Kajamies, A., Vauras, M., & Kinnunen, R. (2010). Instructing low-achievers in mathematical word problem solving. *Scandinavian Journal of Educational Research, 54*(4), 335–355.

Kajamies, A., Vauras, M., Kinnunen, R., & Iiskala, T. (2003). *MATTE—Matematiikan sanallisten tehtävien ratkaisutaidon ja laskutaidon arviointi [Evaluation of the student's mathematical word problem solving and arithmetic skills]*. Turku: University of Turku, Centre for Learning Research.

Kingsdorf, S., & Krawec, J. (2014). Error analysis of mathematical word problem solving across students with and without learning disabilities. *Learning Disabilities Research & Practice, 29*(2), 66–74. https://doi.org/10.1111/ldrp.12029.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *The Psychological Review, 92*(1), 109–129. https://doi.org/10.1037/0033-295X.92.1.109.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: effects of representations on quantitative reasoning. *Journal of the Learning Sciences., 13*(2), 129–164. https://doi.org/10.1207/s15327809jls1302_1.

LeBlanc, M. D., & Weber-Russell, S. (1996). Text integration and mathematical connections: A computer model of arithmetic word problem solving. *Cognitive Science., 20*(3), 357–407. https://doi.org/10.1207/s15516709cog2003_2.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21*(1), 83–90. https://doi.org/10.1007/BF00311017.

Lindeman, J. (1998). *Ala-asteen lukutesti [Comprehensive school reading test; in Finnish]*. Turku: University of Turku, Centre for Learning Research.

Lobato, J. (2012). The actor-oriented transfer perspective and its contributions to educational research and practice. *Educational Psychologist, 47*(3), 232–247.

Mason, J. (2001). Modelling modeling: Where is the centre of gravity of-for-when teaching modeling? In J. F. Mators, W. Blum, S. K. Houston, & S. P. Carreira (Eds.), *Modelling and mathematics education. ICTMA 9: Applications in science and technology* (pp. 36–61). Chichester: Horwood.

Mason, L., & Scrivani, L. (2004). Enhancing students' mathematical beliefs: An intervention study. *Learning and Instruction, 14*(2), 153–176. https://doi.org/10.1016/j.learninstruc.2004.01.002.

Montague, M., Krawec, J., Enders, C., & Dietz, S. (2014). The effects of cognitive strategy instruction on math problem solving of middle-school students of varying ability. *Journal of Educational Psychology, 106*(2), 469–481. https://doi.org/10.1037/a0035176.

Niss, M. (2015). Prescriptive modelling—challenges and opportunities. In G. A. Stillman, W. Blum, & M. S. Biembengut (Eds.), *Mathematical modelling in education research and practice, international perspectives on the teaching and learning of mathematical modelling* (pp. 67–79). Cham: Springer.

Parkash, V., & Kumar, U. (2016). Personality assessment paradigms. In U. Kumar (Ed.), *The Wiley handbook of personality assessment* (pp. 302–322). Hoboken: Wiley-Blackwell.

Pongsakdi, N., Laakkonen, E., Laine, T., Veermans, K., Hannula-Sormunen, M., & Lehtinen, E. (2019). The role of beliefs and motivational variables in enhancing word problem solving. *Scandinavian Journal of Educational Research, 63*(2), 179–197. https://doi.org/10.1080/00313831.2017.1336475.

Pongsakdi, N., Laine, T., Veermans, K., Hannula-Sormunen, M., & Lehtinen, E. (2016). Improving word problem performance in elementary school students by enriching word problems used in

mathematics teaching. *Nordic Studies in Mathematics Education, 21*(2), 23–44.

Quintero, A. H. (1983). Conceptual understanding in solving two-step word problems with a ratio. *Journal for Research in Mathematics Education, 14*(2), 102–112. https://doi.org/10.2307/748578.

Räsänen, P. (2004). *RMAT: Laskutaidon testi 9–12 -vuotiaille [RMAT: A test for arithmetical skills for ages 9 to 12; in Finnish]*. Jyväskylä: Niilo Mäki Institute.

Räsänen, P. (1993). Consistency between two mathematical screening tests and teacher evaluations. In *Paper presented at the III European Congress of Psychology*. Tampere, Finland. July.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reusser, K. (1992). From text to situation to equation: Cognitive simulation of understanding and solving mathematical word problems. *Learning and Instruction, 2,* 477–497.

Reusser, K., Kampfer, A., Sprenger, M., Staub, F., Stebler, R., & Stussi, R. (1990) *Tutoring mathematical word problems using solution trees*. Research Report No. 8, Abteilung Pädagogishe Psychologie, Universität Bern, Switzerland.

Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of language about quantities and of solving problems. *Cognition & Instruction, 5*(1), 49–101. https://doi.org/10.1207/s1532690xci0501_2.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*(5), 1–25.

Stage, C. (2003). *Classical test theory or item response theory: The Swedish experience (No. 42)*. Umeå: Kluwer Academic Publisher.

Swanson, H. L., Cooney, J. B., & Brock, S. (1993). The influence of working memory and classification ability on children's word problem solution. *Journal of Experimental Child Psychology, 55*(3), 374–395. https://doi.org/10.1006/jecp.1993.1021.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Verhelst, N. (2002). *Testing the unidimensionality assumption of the Rasch model (Measurement and Research Department Reports, 2002-1)*. Arnhem: CITO.

Verschaffel, L., & De Corte, E. (1997). Teaching realistic mathematical modeling in the elementary school: A teaching experiment with fifth graders. *Journal for Research in Mathematics Education, 28*(5), 577–601.

Verschaffel, L., Depaepe, F., & Van Dooren, W. (2015). Individual differences in word problem solving. In R. C. Kadosh & A. Dowker (Eds.), *Oxford handbook of numerical cognition* (pp. 953–974). Oxford: Oxford University Press.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Lisse: Swets and Zeitlinger.

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J. E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology, 28*(4), 409–426. https://doi.org/10.1080/01443410701708228.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.