Identifying obstacles is not enough for everybody: Differential efficacy of an intervention fostering fifth graders’ comprehension for word problems

Jennifer Dröse & Susanne Prediger

TU Dortmund University

Source. This is the web version with the accepted manuscript Dröse, Jennifer & Prediger, Susanne (2021). Identifying obstacles is not enough for everybody – Differential efficacy of an intervention fostering fifth graders’ comprehension for word problems. Studies in Educational Evaluation, 68(100953), 1–15. doi.org/10.1016/j.stueduc.2020.100953. The web version is published under for CC-license BY-NC-ND. (Share-it link for open access https://authors.elsevier.com/c/1cEoH4Mg0Ws11)

Abstract. For students’ success in treating mathematical word problems, various language obstacles have been identified. Interventions are required to teach students how to overcome these obstacles, as well as professional development and curriculum support for teachers to implement them. But how much support do teachers need? In a controlled trial evaluation study, two interventions for n = 275 fifth graders are compared: The language-responsive intervention uses the variation principle and the principle of strategic scaffolding to foster comprehension strategies and language awareness. The information-problems-only intervention uses the same set of word problems, but teachers were only informed about typical obstacles. For the evaluation study, the ANOVA shows that in both interventions, students have significantly more learning gains in their word problem comprehension than in the no-treatment group, and substantially more in the language-responsive intervention than in the information-problems-only intervention. Students with different language backgrounds profit comparably from the language-responsive interventions, but in different subscales. In conclusion, identifying obstacles is not enough for implementing effective teaching, it should be enhanced by curriculum support for fostering approaches.

Keywords. Fostering comprehension strategies, fostering language awareness, mathematical word problems, noticing and interpreting language syntactic features

0. Introduction

How can innovations be developed and implemented that help students to overcome typical challenges? We focus this major question on word problems, a typical type of tasks that play an important role in mathematics classrooms as they can enhance or assess reality-oriented considerations. The process of reading, understanding and solving mathematical word problems has widely been described, e.g. by Reusser (1997) for arithmetic word problems. Empirical studies contributed to identifying students’ challenges with word problem in various obstacles (Reusser, 1997; Verschaffel, Greer, & de Corte, 2000). Obstacles, defined as potentially challenging features, were shown to occur in various dimensions, among them mathematical, semantical, structural, strategic, conceptual and language obstacles (Reusser, 1997; Verschaffel et al., 2000; Daroczy, Wolska, Meurers, & Nuerk, 2015). As a consequence of identifying obstacles, effective instructional approaches have been developed and evaluated for mastering conceptual obstacles in word problems (e.g., Van Dooren, De Bock, Hessels, Janssens, & Verschaffel, 2004;), for mastering mathematical, structural and semantical obstacles (Xin & Jitendra, 1999; Zheng, Flynn, & Swanson, 2013; Rudnitsky, Etheredge, Freeman, & Gilbert, 1995), and also for strategic and/or habitual obstacles (e.g.,
Reusser, 1997; Verschaffel, Greer, Dooren, & Mukhopadhyay, 2010; Prediger & Krägeloh, 2015).

This article focuses on language obstacles which comprise e.g. unfamiliar vocabulary or specific meanings of words (e.g. of prepositions or technical terms) on the word level or passive voice and complexe subclauses on the sentence level (Daroczy et al. 2015). So far, empirical findings on language obstacles have been mainly used for avoiding language biases in assessments for students with limited academic language proficiency (e.g. Abedi & Lord, 2001; Haag, Heppt, Roppelt, & Stanat, 2015). Although this reductionist approach is important for fair assessments, it is not sufficient for learning situations, as already Austin and Howson (1979) have argued: “An alternative to simplifying textual material is that of providing children with special reading instruction intended to help them approach and assimilate mathematical vocabulary and phraseology with greater confidence and facility” (p. 173). The call for amplifying approaches has resulted in various approaches for enhancing students’ comprehension strategies (see above), but not many approaches focus on language obstacles on a structural level (as summarized by Solano-Flores, 2010; Erath, Ingram, Moschkovich, & Prediger, 2021). Thus, limited empirical evidence exists so far for the efficacy of instructional approaches for fostering students’ competency in overcoming language obstacles.

This article intends to reduce this research gap by investigating the efficacy and implementability of a language-responsive intervention focusing on fifth graders’ strategies and language awareness. The designed teaching-learning arrangement aims at fostering reading and understanding of word problems on two levels of reading processes: (1) on the text level, i.e. fostering subject-specific reading and understanding strategies through strategic scaffolding with concept maps as a strategic scaffold (Hannafin, Land, & Oliver, 1999) and (2) on the sentence level, i.e. fostering syntactic language awareness through the variation principle (Marton & Pang, 2006).

Whenever innovations are implemented in classroom practice, the big question is how much support teachers need for implementing the innovation (Pellegrini, Lake, Inns, & Slavin, 2018; Saxe, Gearhart, & Nasir, 2001). Is it enough to inform the teachers about typical obstacles, or do they also need support in realizing fostering approaches? In our evaluation study, the language-responsive intervention is compared to an information-problems-only intervention (in which teachers were informed about typical language obstacles, but not supported to foster their students in overcoming them), and a no-treatment control group. The teachers’ preparation for both intervention groups is in line with teachers’ demands for the possibilities of direct implementation of PD themes and data into their teaching practice (Wayman & Jimerson, 2013).

The following section introduces the empirical and theoretical background on typical language obstacles, and specifies comprehension strategies and language awareness required for overcoming them. This is the base for justifying the design principles underlying the language-responsive intervention presented in the second section. Its efficacy is investigated by a cluster-randomized controlled trial for which the research design of the evaluation study and the methods are explicated in the third section. The empirical section then presents the differential results for students with high and low general reading proficiency of the study.
1. **Theoretical and empirical backgrounds on language obstacles in word problems**

Empirical research on students’ processes of solving arithmetic word problems have identified crucial steps and located typical obstacles in these steps (Reusser, 1997; Plath & Leiß, 2018):

1. **Understanding of the textbase** includes reading competencies on a basic level. Potentially, language obstacles but also habitual obstacles can occur in this step (Prediger, Wilhelm, Büchter, Gürsoy & Benholz, 2018).

2. **Understanding of the situation** requires generating a mental model of the situation model and a problem model. Here, obstacles in context knowledge as well as language and strategic obstacles can be located (Nesher & Teubal, 1975; Verschaffel et al., 2000). This has been lined out to be one of the essential steps in word problem solving (Plath & Leiss, 2018).

3. **Mathematization** is important to form a mathematical problem model and reduce it to the connective structure or solution equation. Conceptual obstacles can often be located in this step.

4. **Calculation** for finding the numerical answer. Here, calculation errors can occur.

5. **Interpretation of the answer** to formulate a situational answer. Context knowledge is needed for interpreting numerical answers.

Whereas various obstacles can occur across the five steps, language obstacles, focused in this article, concern the first and second step, mainly the reading of the text and the construction of the situation model (Plath & Leiß, 2018; Leiß, Plath, & Schwippert, 2019). The situation model is denoted as “the personal cognitive structure to which the process of understanding is directed. The situation model is the cognitive correlate to the situation structure either supposed from the author’s point of view or understood from the reader’s point of view” (Reusser, 1989, p.136f. translated by Plath & Leiss, 2018, p.161).

In the following, we present textual features identified as potential language obstacles (in the first subsection) and students’ competencies for overcoming them (in the second subsection).

**1.1 Textual features as potential language obstacles in word problems**

Language obstacles in word problems have been empirically identified in large scale assessments (by analyzing differential item functioning for students with low and high language proficiency, e.g., Abedi & Lord, 2001; Dyrvold, Bergqvist, & Österholm, 2015; Haag, Heppt, Stanat, Kuhl, & Pant, 2013; Prediger, Wilhelm, Büchter, Gürsoy, & Benholz, 2018), in cognitive lab studies (by qualitatively analyzing students’ solving processes, e.g., Prediger et al., 2018) and in simplification studies (by comparing empirical item difficulty for original and simplified items, e.g. Abedi & Lord, 2001; Haag et al., 2015). In their simplification studies, Abedi & Lord (2001) simplified linguistic features in order to familiarity of vocabulary, voice of verb phrase, length of words, conditional and relative clauses as well as questions phrases and impersonal expressions.
Haag et al. (2015) took up those features and simplified academic language complexity by reduction of lexical and grammatical academic language features that have been construct-irrelevant, e.g. long words, compounds, sentence length and passive voice. The results depend on the age of students, types of items and positioning of the obstacles in the text (Daroczy et al., 2015). Repeatedly identified obstacles comprise all levels of language:

- **the word level** comprises unfamiliar vocabulary or specific meanings of words (e.g. of prepositions or technical terms), their flexion (e.g. comparatives or grammatical cases) and composition (e.g. compounds);
- **the sentence level** comprises syntactical complexity of phrases and clauses (e.g. nested prepositional phrases or complex subclauses, passive voice or unusual subject-object orders (in German) and unexpected references by pronouns or condensed phrases).
- **the text level** comprises all features that effect the texts cohesion and its thema-rhema-structure.

The different levels are exemplified by the word problem in Figure 1. Some of those language obstacles have been considered as unnecessary difficulties or construct-irrelevant variances (Abedi & Lord, 2001). In contrast, the current study focuses on language obstacles that are construct-relevant and repeatedly occurring in mathematical language, as displayed in Figure 1 and elaborated below.

The problem in Figure 1 poses major challenges on the text level as constructing the situation model requires creating cohesion between several information and organizing them in four arithmetical operations either in one expression or a multi-step calculation. Potential obstacles on the word level might comprise the unfamiliar word “penguin compound”, obstacles on the sentence level might be located in the passive voice “will be moved”. In Version B, an additional obstacle on the sentence level occurs in the condensed phrase “5 fishes more” which substitutes an unexplicated comparative phrase (“5 fishes more than the chicks”) and therefore poses challenges on the text level for identifying the references in the text.

Although solid empirical findings exist on typical language obstacles, simplifying studies only revealed moderate effects gains in students’ performance by eliminating specific obstacles (Abedi & Lord, 2001; Haag et al., 2015). This calls for research that contributes to fostering
students in overcoming the language obstacles (Erath et al., 2021). Adopting such an amplifying rather than a purely reductionist approach as requested by Austin and Howson (1979) requires further research for specifying the relevant components in students’ competence for mastering language demands, as the process of comprehending a text and constructing a situation model is not only influenced by features of the text, but also by readers’ competence, here in word problem comprehension (Leiß et al., 2019).

1.2 Students’ strategies and language awareness for successfully overcoming language obstacles

Adapting Willenbergs (2007) definitions of general reading comprehension in a topic-specific way, mathematical word problem comprehension is defined as the disposition of a person to understand mathematical aspects while reading word problems and constructing situation models (Dröse, 2019). We focus especially on two components of word problem comprehension that occur on different levels of the reading and understanding process: (a) reading and comprehension strategies on text level and (b) language awareness on sentence level. Reading and understanding on sentence and text level constitute hierarchically higher levels of reading, and both components are needed as students’ competencies for mathematic-specific reading comprehension (Dröse, 2019).

(a) Comprehension strategies

Comprehension strategies are the component of word problem comprehension, which has been most often investigated and fostered in mathematics education research (Capraro, Capraro, & Rupley, 2012). Strategies are defined as goal-oriented mental actions which steer cognitive or metacognitive processes (ibid.).

In the 1970ies, the empirical research on comprehension strategies started with identifying students’ inappropriate strategies such as superficial key word strategies or other direct translation strategies promoting hasty jumps to the mathematization without constructing an appropriate situation model (Nesher & Teubal, 1975; Newman, 1977; Carpenter, Kepner, Corbitt, Lindquist, & Reys, 1980).

Taking into consideration that some studies identified general reading comprehension as predictor for the success of solving mathematical word problems (Bos et al., 2012), different studies aimed at investigating whether fostering general reading comprehension has an impact on word problem solving. In studies of Hellmich and Förder (2015) for primary school as well as Hagen, Leiß & Schwippert (2017) for secondary school, fostering general reading comprehension has been shown not to enhance students’ mathematics-specific reading of word problems. These findings support our assumption that reading and comprehension strategies have to be investigated mathematics-specifically.

For specifying the possible learning content for successful comprehension strategies, cognitive psychologists identified subject-specific strategies used by successful word problem solvers (Reusser, 1994; Hegarty, Mayer, & Monk, 1995). For the age group in view of this article (Grade 5/6), three main strategies have been identified as most relevant (Prediger & Krägeloh, 2015):
(S1) focus on the question and find relevant information concerning the question (Reusser, 1994)

(S2) focus on the meaning of the relevant information (Reusser, 1994; Capraro et al., 2012); and

(S3) focus on relationships among relevant information (Prediger & Krägeloh, 2015; called relational processing in Hegarty et al., 1995; similarly in Kintsch & Greeno, 1985; Reusser, 1994).

These strategies are called mathematic-specific strategies in this study, without excluding that they might also be useful in other subjects (this should be investigated subject-specifically). Meanwhile, mathematic-specific reading strategies was accepted as relevant part of mathematic-specific reading comprehension (Dröse, 2019; Reusser, 1994). As the relation between general reading comprehension strategies and mathematics-specific strategies is still not sufficiently understood (Hagena et al, 2017), the presented study focuses on mathematics-specific reading and comprehension strategies.

(b) Syntactic language awareness

These comprehension strategies can steer the individuals’ processes of reading and constructing an appropriate situation model. For word problems with language obstacles on the word and sentence level, however, the strategies must be complemented by a second component. For instance, the word “more” in the phrase “5 fishes more” of the Penguin Problem Version B in Figure 1 is not even noticed by many students (Dröse, 2019). If they notice it so, it is not automatically interpreted as substituting the phrase “5 fishes more than the chicks”. The competency component underlying these processes of noticing and interpreting language subtleties is here conceptualized as syntactic language awareness (Dröse & Prediger, in press).

Language awareness is a widely considered competence construct in language education research (Clark, 1978; James & Garrett, 1992; Gracia, 2017; Wildemann, Akbulut, & Bien-Miller, 2016). It is defined broadly as the “explicit knowledge about language, and conscious perception and sensitivity in […] language use” (Association for Language Awareness Homepage, quoted from Gracia, 2017, p. 264).

For our research context, we refine the general conceptualization to syntactic language awareness by referring it to language features mainly on the sentence level. This adaptation is needed, as it adjusts language awareness to the specific needs in word problem solving, namely grasping possible subtle syntactic features for understanding the propositional structure of the word problem in order to generate a mental model of the situation (Dröse & Prediger, in press).

To substantiate syntactic language awareness, we draw upon Bialystok’s and Ryan’s (1985) distinction of two cognitive dimensions (analyzed knowledge & cognitive control), and their specification as learners’ skills: the process of noticing language subtleties requires the skill of “control of attentional procedures to select and process specific linguistic information” (Bialystok, 1986, p. 498). Once having noticed the features, students must interpret them that means, translate to their meanings in the situation model and the mathematical structures. Summing up (Dröse & Prediger, in press): Language awareness is expressed in two cognitive processes:
• **noticing** involves the experience that small language details can matter and the sensitivity to perceive them in the word problem text;
• **interpreting** describes the process of decoding the meaning of language details for the text’s situation model by relating the noticed language feature to previous language knowledge (e.g. grammatical knowledge) and therefore anticipating its’ meaning (Dröse, 2019, p. 57).

Borrowing the theoretical construct of language awareness for describing this component of word problem comprehension allows the researchers to transfer also design principles for fostering them, as will be explained in the next section. Although the processes of noticing and interpreting are subject-independent processes, their concretization has to be done subject-specifically, namely: Which language features have to be noticed and how can they be interpreted for the underlying situation of the word problem? (Dröse & Prediger, in press).

Furthermore, noticing and interpreting language features might also be connected to general reading competency. General reading competency is defined as a graded disposition of a reader and the willingness to use it goal orientated (Willenberg, 2007; Gold, 2010). The processes of noticing and interpreting might be related to this general reading competency, but they are not always used consciously and therefore not always used in goal-oriented ways.

Additionally, noticing and interpreting language features might be influenced also by multilingual backgrounds: as typical language obstacles differ between different languages, students with other family languages might have different experiences in noticing and interpreting those language obstacles (see Haag et al., 2013, for connection of multilingual background and general mathematical proficiency). As interpreting syntactic features requires explicit or implicit knowledge about language features, it is an interesting research question in how far in particular multilingual students profit from sensitizing them (Dröse, 2019).

Therefore, the learners’ general reading competency and the multilingual background should be taken into account and studied separately in intervention studies.

2. **Language-responsive design principles for fostering comprehension strategies and language awareness for mathematical word problems**

As Heid et al. (2006) have emphasized in their strategic position paper, research should not stop with identifying obstacles and specifying learning contents, i.e. components of word problem comprehension, but should bridge the gap to classroom practices by collaboratively developing instructional approaches and concrete teaching units for fostering the learning contents. For language in mathematics classrooms, various teaching approaches are summarized by Erath et al. (2021).

With respect to mathematical word problems, instructional approaches were developed and tested for mathematics-specific comprehension strategies (e.g., Reusser, 1994; Schukajlow, Kolter, & Blum, 2015), but not yet for language awareness.

While drawing upon approaches for fostering comprehension strategies by strategical scaffolding (to be presented in the first subsection), this paper intends to show that also language
awareness as the second component can be effectively fostered. For this, we transfer the variation principle to fostering language awareness for word problems (in the second subsection). For both design principles, the realization in a language-responsive teaching unit is illustrated.

2.1 Strategic scaffolding as a design principle for fostering comprehension strategies

For fostering comprehension strategies, the design principle of strategic scaffolding (described by Hannafin et al., 1999), has proven to be effective. In general, the scaffolding principle aims at supporting students in reaching the zone of proximal development (Wood, Bruner, & Ross, 1976, p. 90). That means the zone of tasks that at the moment cannot be reached by the learner’s state of development, but might be reached by the learner alone after a sufficient period of support (Wood et al., 1976). The scaffolding principle was explored in multiple research traditions. It draws upon the idea that a more skilled or knowledgeable person shares work with the learner in order to make the learner solve tasks that he or she could not have solved alone (Reiser & Tabak, 2017). For this, scaffolds can be used as long as they are needed and should be faded out in the end (Lajoie, 2005; Hannafin et al., 1999). Scaffolding approaches can be classified according to what, how and through what something is scaffolded (Lajoie, 2005). Focusing on the aim of scaffolding, conceptual, procedural, metacognitive and strategic scaffolding can be distinguished from one another (Hannafin et al., 1999).

Strategic scaffolding especially “supports analysis, planning, strategy, and tactical decisions during open-ended learning. It focuses on approaches for identifying and selecting needed information, evaluating available resources, and relating new knowledge to existing knowledge and experience” (Hannafin et al., 1999, p. 133).

Several instructional approaches exist that offer temporary guidance for students’ comprehension processes, even if not named strategic scaffolding by their designers: for example, prompts for reading steps such as “ask questions to the text” or for problem solving steps (e.g. Mevarech, Terkieltaub, Vinberger, & Nevet, 2010), or solution plans for modelling steps (e.g. Schukajlow et al., 2015).

A more graphical strategical scaffold was suggested by Reusser (1994) in the computer-assisted tool Heron (example in Figure 2). This scaffold has the potential to scaffold the comprehension strategy S1 and S2 (see above): It helps to focus on the relevant information together with their meanings. But the necessary focus on relations between the information (Strategy S3) is immediately connected to the mathematical operationalization which turned out to be too rigid for many students (Prediger & Krägeloh, 2015).

![Example word problem: Eve's Scarf](image)

Eve has promised to knit a scarf for her grandfather’s 70th birthday. Because she spends a lot of time drawing, she has only finished 45 rows so far. Grandfather’s birthday will be in 15 days. Eve is afraid that the scarf will not be ready in time. The complete scarf should be 165 rows long. Her mother has promised to help her. She knits 5 rows every time Eve has done 15 by herself. Now Eve wants to know how many rows she has to knit every day in order to finish the scarf in time.

![Figure 2. Solution tree in computer assisted tool HERON](image)

(Reusser, 1994, p. 179)
Combining the advantages of solution plans with those of graphical scaffolds became possible by using concept maps such as in Figure 3 for the Penguin Problem from Figure 1 (from Dröse, Prediger & Marcus, 2017; Dröse & Prediger, in press).

Concept maps have proven to be a useful tool in different instructional approaches, displayed in Afamasaga-Fuata’I (2009), but have rarely been used to foster students’ comprehension strategies. Meanwhile, external graphical representations might be supportive as they “may facilitate cognitive representation of the information” (p. 417). They allow the processing of information by reducing, organizing and elaborating them as well as filling in gaps between information and infer new information (Nesbit & Adesope, 2006).

The concept map is introduced together with the solution plan printed in Figure 3 which regulates the steps of producing the concept map. Three main steps resonates with the identified strategies:

(2) Write the question card. This supports S1: Focus on the question.

(3) Write the information card. This can be numbers with unit and explanation or other aspects of the word problem. This supports S2: Focus on the relevant information together with their meaning.

(4) Connect information cards and display the relationship by other information cards. This supports S3: Focus on the relationships between the relevant information.

In order to introduce (general) comprehension strategies, researchers in language education have proposed a sequence of steps (Yelland & Masters, 2007; Puntambekar & Hübscher, 2005). We have adapted these steps for introducing mathematics-specific comprehension strategies by the teachers:

**Step 1.** Students get to know the strategies named above (S1, S2 and S3) and the strategic scaffold

**Step 2.** Students practice and automatize the use of comprehension strategies (S1, S2 and D3) through the strategic scaffold
Step 3. Students use the comprehension strategies independently through the strategic scaffold without needing support of teacher or fellow students to master the comprehension of the given word problems

Step 4. Fading-out of the strategic scaffold takes place in students’ use of it mostly initiated by the teacher

2.2 Variation as an imported design principle for fostering language awareness

For fostering syntactic language awareness of language subtleties, the design principle of variation is used. This design principle is already well established and known as the principle of syntactic contrast in the field of language education (Melzer, 2013), where the concept of language awareness stems from. By using the syntactic contrast, texts with differences in language subtleties are systematically contrasted to discuss those structures (Melzer, 2013). So far, the principle of syntactic contrast focus mainly on linguistic aspects and does not take subject-specific features into consideration.

A similar approach, known as the variation principle, can also be found in the field of mathematics education in line with the Chinese tradition of Bianshi teaching (Pang, Bao, & Ki, 2017) and the European tradition of Variation Theory (Marton & Pang, 2006). The principle’s basic assumption is that students become aware and can distinguish essential from non-essential features of a learning content if those features are systematically varied (Gu, Huang, & Gu, 2017). Three main characteristics are essential for the variation principle for language obstacles in word problems (Dröse & Prediger, in press; Dröse, 2019):

(C1) The varied tasks or word problems are systematically comparable to each other. This means, that there are systematic changes and invariants. Comparing the tasks aims at identifying those changes and invariants (Pang et al., 2017).

(C2) Variation of sentence features within one word problem. Following the tradition of the Variation Theory (Marton & Pang, 2006) and the syntactical contrast (Melzer, 2013) essential features are systematically varied, cf. figure 1 for a variation of the penguin problem, to focus the students’ attention on those features.

(C3) Invariance of syntactic features within a sequence of tasks. Following the tradition of Bianshi teaching the relevant features are kept the same within a sequence of word problems to allow generalization against the background of changing other features of the word problem, e.g. context, numbers or mathematical operations (Pang et al., 2017). Figure 4 shows the next task “Taking photographs in the zoo” following the “penguin” task. The essential feature is still substituting a comparative phrase but embedded in a different task.
For fostering syntactic language awareness for language obstacles in word problems, three characteristics of the tasks are included and sequenced in the teaching unit as follows (in Dröse, Prediger, & Marcus, 2017):

Phase 1: Getting to know noticing and interpreting for word problems with linear structure
   i. Introduction of noticing and interpreting of syntactic features (C1 & C2)
   ii. Practice of noticing and interpreting of syntactic features (C2 & C3)

Phase 2: Getting to know noticing and interpreting for word problems with non-linear structure and varied subject-object position
   i. Introduction of noticing and interpreting of syntactic features (C1 & C2)
   ii. Practice of noticing and interpreting of syntactic features (C2 & C3)

Phase 3: Getting to know noticing and interpreting for word problems with non-linear structure and varied substitution of a comparative phrase
   i. Introduction of noticing and interpreting of syntactic features (C1 & C2)
   ii. Practice of noticing and interpreting of syntactic features (C2 & C3)

The variation principle and the principle of strategic scaffolding form the two main design principle of the teaching units. Previous qualitative analysis of the empirical data gives hints at the situational effects of the teaching unit (Dröse, 2019; Dröse & Prediger, in press). Meanwhile due to the limited number of cases the qualitative data analysis was not able to give sufficient evidence of differences among learners with different language background that means multilingual background and general reading competency. Therefore, the effects of the design principles and the intervention as a whole should be investigated taking also differential analysis of students’ language background into consideration.

**Figure 4.** Typical obstacle in a word problem: substituting a comparative phrase “2 less than the starfish” by “2 less” (from Dröse, Prediger, & Marcus, 2017)
2.3 Research questions for the current study

Heid et al. (2006) criticized the analytical research that solely identifies students’ obstacles and called for more practice-near research that supports teaching practices for overcoming the obstacles in classrooms. Within the last decades, many design research approaches have followed this request and developed instructional approaches. However, there is still a need to study which support teachers need to implement innovative instructional approaches (Pellegrini, Lake, Inns, & Slavin, 2018; Saxe, Gearhart, & Nasir, 2001).

For instruction on general reading comprehension, Förster and Souvignier (2015) could show that it is not sufficient to just inform teachers about assessed student abilities. However, the effects of supporting teachers through teaching-learning-arrangements based on design principles have rarely been measured in direct comparison to approaches of simply informing teachers about obstacles. This is investigated in the current study with the following research question:

**RQ1.** To what extend do students’ learning gains for comprehension of word problems differ between the language-responsive intervention, the information-problems-only intervention and the no-treatment group?

Language-responsive instructional approaches have originally been developed with respect to multilingual students and students with low reading proficiency (e.g., Gibbons, 2002). For applying them in mainstream classrooms, it is important to consider whether (1) reading proficiency and (2) multilingual background really impacts the students’ learning gains of word problem comprehension as a whole and for its different components.

As described in the previous section, the background factor reading proficiency might be important as some studies identified general reading comprehension as a predictor for the success to solve mathematical word problems (Bos et al., 2012), but the relationship between general reading comprehension and mathematic-specific reading strategies has rarely been investigated.

In addition, the background factor multilingualism might be relevant as typical language obstacles differ between different languages, so students with other family languages might have different experiences in noticing and interpreting those language obstacles (see Haag et al., 2013, for connection of multilingual background and general mathematical proficiency) and especially interpreting syntactic features requires explicit or implicit knowledge about language features, which might not be familiar to some multilingual students (Dröse, 2019).

In the following, multilingual background and reading proficiency will be summarized as language background factors for studying differential aptitudes and aptitude treatment interactions (Snow, 1991):

**RQ2.** In which way do the language background factors (multilingual background and reading proficiency) impact the students’ learning gains for comprehension of word problems?

**RQ3.** For which language background factors are there differences in students’ starting points for strategy use and language awareness (noticing and interpreting language obstacles)?
RQ4. For which language background factors are there differences in students’ learning gains for strategy use and language awareness (noticing and interpreting language obstacles)?

Only if all students with their different language backgrounds can profit from an intervention, it can be successfully implemented into mainstream classrooms with their usual language diversity. Therefore, students’ differential aptitudes and differential learning gains are important to study for investigating implementability.

3. Methods of the controlled field trial

3.1 Research design and intervention forms

Overview. Figure 5 gives an overview of the design of the cluster-randomized controlled trial with pre- and post-test with n = 275 students and n = 22 teachers of middle schools in urban areas. Two competing interventions served as the independent variable compared to a no-treatment control group. Students’ comprehension of word problems and especially their syntactic language awareness were the dependent variables. Language background (multilingual background and reading proficiency) were considered as control variables and as focus of attention for differential analyses. Additionally, age, gender, and socio-economic status were considered as control variables.

![Figure 5. Design and Measures of the cluster-randomized controlled trial with pre- and post-test design](image-url)
**Intervention forms as independent variable.** The classroom interventions were taught by the classes’ regular teachers with comparable backgrounds and comprised 5-6 lessons of 90 min each. Two forms of interventions and a waiting control group were compared. The two interventions differ in PD content and the used teaching unit. Comparable settings have already been investigated for other subjects to a small extent (e.g. cf. ter Beek, Opdenakker, Deunk, Strijbos 2019 for reading strategy training in history lessons):

**(LRI)** The students in the **language-responsive intervention** (abbreviated LRI) were taught the teaching unit based on the design principles of strategic scaffolding by concept maps and solution plan for fostering students’ comprehension strategies and the variation principle for fostering syntactic language awareness for word problems as described in Section 2. The worksheets and manual of the teaching units are published in Dröse, Prediger, & Marcus (2017). The teachers were prepared in a *preparation workshop* (of 90 minutes). The preparation workshop was delivered by members of the research team. The teachers worked out typical strategic and language obstacles with the facilitators from students’ cases and example word problems that were presented to them (approx. 40 min). Afterwards they were introduced to the principles of strategic scaffolding and the variation principle, and to their implementation in the teaching material. Then, they tried out the design principles and material themselves and clarified their implementation with the facilitators (approx. 50 min). The teachers participated actively in the elaboration of knowledge on obstacles and design principles as well as in getting to know the teaching-learning arrangement. They were encouraged to ask clarifying questions and cooperate at all time.

**(IPI)** The students in the **information-problems-only intervention** (abbreviated IPI) worked on a *problems-only-teaching unit*, with the same set of word problems that means identical word problem texts as in LRI provided for the teacher by the research group, but without strategic scaffolds that means no concept maps, no solution plan and without varied problems. The teachers were prepared in an *information-only-workshop* (of 90 minutes). The preparation workshop was delivered by members of the research team. The teachers worked out typical strategic and language obstacles with the facilitators from students’ cases and example word problems that were presented to them (approx. 40 min.). They received detailed additional information about the obstacles and the used word problems of the teaching-learning arrangement (approx. 20 min.). The teachers were not introduced to the principles of strategic scaffolding and variation. They were free to choose their instructional approaches for the *problems-only-teaching unit*. They discussed possible approaches with each other in the workshop (approx. 30 min.). The teachers participated actively in the elaboration of knowledge on obstacles. They were encouraged to ask clarifying questions and cooperate at all time.

**(NTG)** The students in the **no-treatment group** (abbreviated NTG) received no-treatment on word problems. Instead they participated in their regular mathematics classrooms with the regular textbook approach not focused on word problems. The teachers did not get any workshop on dealing with word problems. This treatment condition was added for
a baseline testing, ensuring that the pure repetition of test and students natural growth can also be taken into account.

Treatment control was conducted in all groups by collecting and analysing the students’ materials after the teaching unit and by interviews with both groups of teachers. It revealed that both intervention groups realized their teaching units with sufficient fidelity, which means over 80% workbook completion rate, use of concept maps in the language-responsive intervention group.

3.2 Measures for dependent variable and control variables

Dependent variable word problem comprehension with three subscales

The particular learning content in view, students’ word problem comprehension, was assessed in the pre-test and post-test before and after the intervention. The parallel tests had been developed particularly for the project (Dröse, 2019) to be able to assess the use of mathematic-specific reading strategies and the processes of noticing and interpreting. This is necessary in order to capture the depth of the learning contents in the teaching-learning arrangement by a pre- and post-test. As no other written tests exists that capture this particular learning content, the construct validity was assessed by triangulating the test scores with oral interview data (Dröse, 2019).

The tests had a maximum score of 15, differed only in numbers and contexts, and contained two subscales:

Subscale strategy use. The test contains four rich word problems, which require the use of strategies S1–S3 for constructing an adequate situation model. Although students can apply the strategies without being aware of them, the strategies are crucial to be applied for solving the problems. By using inadequate strategies, no correct situation model could be generated. Each of the presented subtasks requires especially using one of the requested strategies: (S1) finding relevant information concerning the problem question, (S2) identifying the meaning of the relevant information and processing them, (S3) the relationship between information in multistep word problems and processing them. For example, it is not possible to solve tasks with focus on (S1) finding relevant information concerning the problem question if students do not focus on the question. Meanwhile, in this study the strategies are not investigated individually but all together to investigate the strategy use as a whole. Students’ written solutions were coded and evaluated with respect to the use of the requested strategies. Mathematically correct calculations within the mathematical model were excluding from assessing comprehension. The maximum score in the subscale was 12.

![Table](Lara gets 25 sticker from her aunt. Her brother is given 11 Sticker by her. (Literally: Her brother, gives she.) How many sticker does Lara have?
Syntactic case analysis of German original text:
Version A: Ihrem Bruder gibt sie 11 Sticker ab. Dativ (Object) Nominativ (Subject)
Version B: Ihr Bruder gibt ihr 11 Sticker. Nominativ (Subject) Dativ (Object)
Lara gets 25 sticker from her aunt. Her brother gives her 11 Sticker. How many sticker does Lara have?

![Figure 6](Example item for noticing subtle language features)
**Subscale noticing of language obstacles.** The subscale on noticing syntactic language features is varied according to German cases, which allows in German to change the order of subject and object in a sentence without using passive voice (cf. Dröse, 2019; Dröse & Prediger, in press). Figure 6 shows an example of this scale. The students’ written solutions were coded as follows: (0) no problem is solved, (1) only one problem is solved and the other one is not solved or in the same way, (2) answers to both problems but switched, (3) both problems are solved as required. The maximum score in the scale was 3. With this subscale, it is possible to assessing noticing of syntactic features as the codes (0) and (1) imply that no difference between the word problems has been noticed, while codes (2) and (3) show that differences among the word problems have been noticed. Word problems have been phrased in an easy way to facilitate interpretation and use for mathematization.

**Additional scale interpreting language obstacles.** In addition to the subscales strategy use and noticing, four varied word problems (printed in Figure 7) were developed to form a scale on interpreting relevant language features in word problems, which is not technical part of the word comprehension scale. This additional scale is needed to capture the process of interpreting although this has not been an explicit learning content of the teaching-learning arrangement. By providing the same items with only slight distinctions, noticing was heavily fostered, so that this scale could capture the grammatical knowledge for correct interpretations. For this, students’ written solutions were coded as follows: (0) no problem is solved, (1) one random problem is solved, (2) just the problems containing the same syntactic structure are solved, (3) problems of both syntactic structures are solved but just one step or just two step problems, (4) three out of four problems are solved, (5) all problems are solved. The maximum score in the scale was 5.

The coding of students’ solutions with respect to the mentioned codes was done by two independent coders. Over all scales it reached an interrater reliability of Cohen’s $\kappa = 0.95$.

<table>
<thead>
<tr>
<th>Hanna and Ben are siblings.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Her brother Ben is given 4 Sticker by her.</em></td>
</tr>
<tr>
<td>(Literal translation: <em>Her brother Ben, gives she.</em>)</td>
</tr>
<tr>
<td>How many sticker does Hanna have?</td>
</tr>
</tbody>
</table>

**Syntactic case analysis of German original text:**

<table>
<thead>
<tr>
<th>Male, Object</th>
<th>Female, Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ihrem Bruder Ben gibt sie 11 Sticker ab.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lukas and Klara are siblings.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>His sister Emma is given 5 € by him.</em></td>
</tr>
<tr>
<td>(Literal translation: <em>His sister Emma, gives he.</em>)</td>
</tr>
<tr>
<td>How many Euro does Klara have?</td>
</tr>
</tbody>
</table>

**Syntactic case analysis of German original text:**

<table>
<thead>
<tr>
<th>Male, Subject</th>
<th>Female, Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seiner Nachbarin Emma gibt er 5 € ab.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lea and Felix are friends.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Her friend Felix gives her 3 DVDs.</em></td>
</tr>
<tr>
<td>How many DVDs does Lea have?</td>
</tr>
</tbody>
</table>

**Syntactic case analysis of German original text:**

<table>
<thead>
<tr>
<th>Male, Subject</th>
<th>Female, Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ihr Freund Felix gibt ihr 13 DVDs ab.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jonas and Emma are neighbours.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>His neighbour Emma gives him 10 €.</em></td>
</tr>
<tr>
<td>How many Euro does Emma have?</td>
</tr>
</tbody>
</table>

**Syntactic case analysis of German original text:**

<table>
<thead>
<tr>
<th>Male, Subject</th>
<th>Female, Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seine Nachbarin Emma gibt ihm 10 € ab.</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 7.** Four word problems in the scale interpreting subtle language obstacles: Variation of German cases and sexus for facilitated noticing
Control variables

Furthermore, the following control variables were taken into account due to their often shown influence on mathematics learning with respect to language issues.

- **Age, gender, multilingual background and socio-economic status.** Age, gender and multilingual background were captured by students’ self-report in a questionnaire. For the multilingual background, students were asked to list all languages spoken with their parents. Socio-economic status was captured by the book scale, an economic and reliable instrument ($r = 0.80$, Paulus, 2009).

- **Reading proficiency.** General reading proficiency was controlled as the component of general language proficiency which is closest to mathematical text comprehension. It was assessed by the ELFE-Test, a reading comprehension speed screening which measures reading speed and comprehension in three scales, on the word level (72 items), the sentence level (28 items), and the text level (20 items). All three scales have high internal consistencies (with Cronbach’s $\alpha$ between .92 and .97) and a retest reliability of $r = 0.93$ (Lenhard & Schneider, 2006). The maximum score is 120, the median 60. The ELFE-Test was administered before the intervention in their regular class setting by the research team. The research team supervised the testing procedure and the test evaluation.

### 3.3 Sample

In total, 22 teachers were randomly assigned to the three treatment conditions (two more teachers in the IPI intervention dropped out for personal reasons). The teachers participated voluntarily in the teacher preparation and classroom intervention. According to informal pre-intervention interviews, all have comparable backgrounds with respect to (a) mathematics teacher certificates for middle schools, (b) teaching experience between 3 and 15 years, (c) previous experience in language-responsive teaching.

The LRI group consisted of $n = 118$ students with pre- and post-test results and the IPI group of $n = 41$ students with pre- and post-test results. As the waiting control group (NTG) was not comparable in this cluster-randomized sampling procedure at first glance, each student of the LRI group was matched to a student of the waiting control group with respect to relevant control variables and pre-test scores, the $n = 118$ matched students then formed the no-treatment group NTG. After this matching procedure, students in the three groups did not differ regarding pre-test results, general reading proficiency, multilingual background, and socio-economic status (four t-tests with $p > .05$ each). Significant but small differences (of less than 9 months) occurred in students’ age between the groups (with $p < .05$ in t-tests).
## 3.4 Hypotheses and methods of quantitative data analysis

In the quantitative data analysis, the following hypothesis was tested for research question RQ1:

(H) The three groups have different learning gains (difference in test scores) in word problem comprehension.

For testing this hypothesis, an analysis of variances with repeated measures was conducted with the software R (Package ez); type-III ANOVA was chosen as it accounts for different group sizes. Differences were measured on a 5% significance level. Intragroup effects were measured by Cohens d.

The research questions RQ2 and RQ4 are more explorative in nature, as hypotheses for advantages could not be derived from literature. For comparing students of different language backgrounds with respect to overall learning gains (RQ2), differential starting points for strategies and language awareness (RQ3) and learning gains in these components (RQ4), we split the sample into dichotomic groups:

(a) Mono vs. Multi: into monolingual and multilingual students (according to their self-reports on language use in their families) and

(b) RP⁺ vs. RP⁻: into students with high and low general reading proficiency. The group of weak readers RP⁻ was formed by a median split for a dichotomic classification with the cut-off-score up to 60 in the ELFE test, strong readers above. This cut-off resonates with the test manual’s distinction between students with reading proficiency below and above average (Lenhard & Schneider, 2006).
4. Results of the differential data analysis

4.1 Significantly higher learning gains in the language-responsive intervention

Research question RQ1 asks for the overall efficacy of the interventions compared to each other. The results in the first lines of Table 2 reveal that both intervention groups had highly significant learning gains ($F_{time/LRI} (1, 117) = 89.00$ with $p < 0.01$ and $\eta^2 = 0.20$; $F_{time/IPI} (1, 40) = 13.86$ with $p < 0.01$ and $\eta^2 = 0.09$). Also the NTG group show significant differences from pre- to post-test ($F_{time/NTG} (1, 117) = 9.77$ with $p < 0.01$), but these baseline gains have very low effect size ($\eta^2 = 0.01$, $d = 0.29$).

Comparing the learning gains reveals that students in LRI developed substantially more comprehension for word problems than in IPI and in NTG ($\Delta m = 2.85$ for LRI, $\Delta m = 2.06$ for IPI and $\Delta m = 0.78$ for NTG), this is also reflected by the effect sizes in Cohen’s $d$ ($d = 0.87$ for LRI, $d = 0.58$ for IPI and $d = 0.29$ for NTG). In addition, the ANOVA confirms significant differences among the three groups ($F_{time \times group} (1, 274) = 13.41$, $p< 0.01$, $\eta^2 = 0.02$). Hence, the results confirm the Hypothesis (H) and tend to show that LRI learned more than IPI than NTG (abbreviated LRI > IPI > NTG).

Table 2. Learning gains in word problem comprehension in all groups

<table>
<thead>
<tr>
<th></th>
<th>Pre-test $m$ (SD)</th>
<th>Post-test $m$ (SD)</th>
<th>Learning gain $\Delta m$ (SD)</th>
<th>Effect size $d$</th>
<th>ANOVA results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LRI</strong> = language-responsive intervention (n = 118)</td>
<td>6.23 (2.48)</td>
<td>9.08 (3.13)</td>
<td>2.85 (3.28)</td>
<td>0.87</td>
<td>$F_{time/LRI} (1, 117) = 89.00$; $p&lt;0.01$, $\eta^2 = 0.20$</td>
</tr>
<tr>
<td><strong>IPI</strong> = information-problems-only intervention (n = 41)</td>
<td>5.62 (2.92)</td>
<td>7.68 (3.59)</td>
<td>2.06 (3.54)</td>
<td>0.58</td>
<td>$F_{time/IPI} (1, 40) = 13.86$, $p&lt;0.01$, $\eta^2 = 0.09$</td>
</tr>
<tr>
<td><strong>NTG</strong> = no-treatment group (n = 118)</td>
<td>6.50 (3.53)</td>
<td>7.27 (3.51)</td>
<td>0.78 (2.69)</td>
<td>0.29</td>
<td>$F_{time/NTG} (1, 117) = 9.77$, $p&lt;0.01$, $\eta^2 = 0.01$</td>
</tr>
</tbody>
</table>

For all groups: $F_{time} (1, 274) = 82.05$, $p < 0.01$, $\eta^2 = 0.07$ and $F_{time \times group} (1, 274) = 13.41$, $p< 0.01$, $\eta^2 = 0.02$

4.2 Differential effects for different language backgrounds

According to research question RQ2, the learning gains were exploratively compared for students with different language backgrounds. Table 3 provides the results for each of these comparisons. Due to small sample size of the subgroups (especially in IPI) the differences among subgroups did not become significant. However, the comparison of the descriptive results and the effect sizes show interesting tendencies:

- Among the monolingual students, the learning gains of NTG/Mono (i.e. monolingual students in no-treatment group) is lowest ($\Delta m = 0.89$, $d = 0.33$), documenting very few learning gains. The learning gains of monolingual students in the LRI group LRI/Mono is substantially higher ($\Delta m = 2.63$, $d = 0.82$) than the learning gains of IPI/Mono ($\Delta m = 1.64$, $d = 0.42$). In conclusion, monolingual students tend to profit more from LRI than from IPI (LRI/Mono > IPI/Mono).
• Among the multilingual students, the learning gains of the group LRI/Multi (i.e. multilingual students in the language-responsive intervention) is highest ($\Delta m = 3.04$, $d = 0.91$), the learning gains of IPI/Multi is lower ($\Delta m = 2.28$, $d = 0.67$) and NTG/Multi show few learning gains ($\Delta m = 0.70$, $d = 0.26$). That means, also multilingual students tend to profit more from LRI than from IPI (LRI/Multi > IPI/Multi), with slightly higher effect sizes than for the monolinguals in both groups.

• For strong readers (i.e., students with high reading proficiency), the intervention LRI has slight advantages ($\Delta m = 2.81$, $d = 0.9$) compared to IPI ($\Delta m = 2.35$, $d = 0.56$). Nevertheless, both intervention groups show substantially higher learning gains that NTG ($\Delta m = 0.41$, $d = 0.14$). We conclude briefly the tendency LRI/RP+ > IPI/RP+.

• Weak readers (students with low reading proficiency) tend to profit substantially more from LRI ($m = 2.88$, $d = 0.90$) than from IPI ($m = 1.79$, $d = 0.62$), briefly written LRI/RP− > IPI/RP−. Meanwhile weak readers profit slightly more from IPI than NTG ($m = 1.13$, $d = 0.42$) concerning learning gains but with smaller effect size.

Summing up, all language groups tend to profit more from the language-responsive intervention than from the other intervention, and remarkably, the multilingual students have the largest difference. These results are worth to be disentangled with respect to the students’ heterogeneous starting points concerning different components of word problem comprehension, namely strategy use and syntactic language awareness (noticing and interpreting language obstacles). These explorations are presented in the following two subsections.

| Table 3. Learning gains in word problem comprehension – comparing students with different language backgrounds |
|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|
| Scores in word problem comprehension          | Pre-test $m$ (SD) | Post-test $m$ (SD) | Learning gains $\Delta m$ (SD) | Effect size $d$ |
| LRI/Mono Monolingual students in LRI (n = 56) | 6.69 (2.79)       | 9.32 (2.69)       | 2.63 (3.22)       | 0.82 |
| IPI/Mono Monolingual students in IPI (n = 14) | 6.21 (3.47)       | 7.86 (3.52)       | 1.64 (3.94)       | 0.42 |
| NTG/Mono Monolingual students in NTG (n = 49) | 6.51 (3.55)       | 7.40 (3.53)       | 0.89 (2.68)       | 0.33 |
| LRI/Multi Multilingual students in LRI (n = 62) | 5.82 (2.09)       | 8.86 (3.49)       | 3.04 (3.34)       | 0.91 |
| IPI/Multi Multilingual students in IPI (n = 27) | 5.31 (2.62)       | 7.59 (3.68)       | 2.28 (3.38)       | 0.67 |
| NTG/Multi Multilingual students in NTG (n = 69) | 6.49 (3.54)       | 7.18 (3.52)       | 0.70 (2.72)       | 0.26 |
| LRI/RP+ Strong readers in LRI (n = 59) | 6.80 (2.71)       | 9.61 (3.25)       | 2.81 (3.21)       | 0.90 |
| IPI/RP+ Strong readers in IPI (n = 20) | 6.22 (3.01)       | 8.57 (3.75)       | 2.35 (4.19)       | 0.56 |
| NTG/RP+ Strong readers in NTG (n = 58) | 8.16 (3.42)       | 8.56 (3.49)       | 0.41 (2.99)       | 0.14 |
| LRI/RP− Weak readers in LRI (n = 59) | 5.67 (2.09)       | 8.55 (2.94)       | 2.88 (3.21)       | 0.90 |
| IPI/RP− Weak readers in IPI (n = 21) | 5.05 (2.79)       | 6.83 (3.30)       | 1.79 (2.87)       | 0.62 |
| NTG/RP− Weak readers in NTG (n = 60) | 4.89 (2.84)       | 6.03 (3.08)       | 1.13 (2.35)       | 0.48 |
4.3 Disentangling differential starting points for students with different language backgrounds into strategy use, noticing and interpreting

For disentangling students’ specific starting points according to research question RQ3, Table 4 shows the different scores in the pre-test for each of the language background groups, differentiated into the scales strategy use, noticing and interpreting language obstacles. The differences were tested for significance by t-tests.

With respect to strategy use in the first lines of Table 4, the first columns show the comparison of monolingual and multilingual students, revealing no significant differences (p = 0.05). In contrast, the strategy use in the pre-test differs highly significantly between students of high and low reading proficiency (p < 0.001) with a high effect size (d = 0.71). These findings allow to generate the hypothesis strategy/RP+ > strategy/RP-, that means, students with high reading proficiency might use more strategies than students with low reading proficiency prior to the intervention, whereas multilingual students and monolingual students seem to be comparable in strategy use.

Table 4. Differences in pre-test scores for scales of word problem comprehension (bold: significant differences)

<table>
<thead>
<tr>
<th>Scales</th>
<th>Monolingual students (n = 119)</th>
<th>Multilingual students (n = 158)</th>
<th>Students with low reading proficiency (n = 140)</th>
<th>Students with high reading proficiency (n = 137)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scores in the pre-test m (SD)</td>
<td>4.63 (2.79)</td>
<td>4.00 (2.56)</td>
<td>3.39 (2.26)</td>
<td>5.18 (2.77)</td>
</tr>
<tr>
<td>Inter group effect size d p</td>
<td>d = 0.24, p = 0.05</td>
<td>d = 0.71, p &lt; 0.001**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noticing of language obstacles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scores in the pre-test m (SD)</td>
<td>1.89 (0.88)</td>
<td>2.05 (0.85)</td>
<td>1.86 (0.87)</td>
<td>2.11 (0.85)</td>
</tr>
<tr>
<td>Inter group effect size d p</td>
<td>d = 0.19, p = 0.12</td>
<td>d = 0.29, p = 0.02 *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpreting of language obstacles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scores in the pre-test m (SD)</td>
<td>3.74 (1.28)</td>
<td>3.27 (1.54)</td>
<td>3.09 (1.56)</td>
<td>3.86 (1.21)</td>
</tr>
<tr>
<td>Inter group effect size d p</td>
<td>d = 0.33, p = 0.01 *</td>
<td>d = 0.55, p &lt; 0.001**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

With respect to noticing of language obstacles, the middle lines of Table 4 show the comparison of monolingual and multilingual students and students of high and low reading proficiency. Comparing monolingual and multilingual students reveals no significant difference (p = 0.12). In contrast, comparing students of high and low reading proficiency reveals significant differences (p < 0.05), with small effect size (d = 0.29). These findings allow to generate the hypothesis noticing/RP+ > noticing/RP-, that means, students with high reading proficiency might notice more syntactic subtleties than students with low reading proficiency prior to the intervention, whereas multilingual students and monolingual students seem to be comparable in noticing.

The lowest lines of Table 4 show the results of the scale interpreting. Significant differences are found between the interpreting of monolingual and multilingual students (p < 0.05, d = 0.33) as well as for students of high and low reading proficiency are highly significant (p < 0.05).
0.001) with higher effect sizes than for strategy use (d = 0.55): the hypotheses interpreting/Mono > interpreting/Multi as well as interpreting/RP+ > interpreting/RP- can be generated. Again, students with high reading proficiency might interpret the syntactic subtleties better than students with low reading proficiency prior to the intervention, and unlike strategy use and noticing, also monolingual students might interpret better than multilingual students.

These findings indicate that noticing syntactic features might be a learning content for all students, independent of their language background. In contrast, interpreting syntactic features seems to be a more relevant learning content especially for multilingual students or students with low reading proficiency.

4.4 Differential effects of the language-responsive intervention on different scales

The slightly heterogeneous starting points for each scale also call for studying the heterogeneous learning gains within each scale, so RQ 4 compares the learning gains for strategy use, noticing and interpreting language obstacles, for the different language groups. Although the ANOVAs showed no significant differences, the comparison of effect sizes can provide interesting tendencies for some components and not for others.

Table 5. Learning gains in strategy use in the language-responsive intervention LRI (adapted from Dröse, 2019, * marks a significant learning gain within the subsample)

<table>
<thead>
<tr>
<th>Scores in strategy use</th>
<th>Pre-test m (SD)</th>
<th>Post-test m (SD)</th>
<th>Learning gain Δ m (SD)</th>
<th>Effect size d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRI</td>
<td>(n = 118)</td>
<td>4.15 (2.22)</td>
<td>6.61 (2.86)</td>
<td>2.46 (3.06)</td>
</tr>
<tr>
<td>LRI_strategy/Mono Monoilinguals in LRI (n = 56)</td>
<td>4.70 (2.51)</td>
<td>6.85 (2.48)</td>
<td>2.15 (3.06)</td>
<td>0.70*</td>
</tr>
<tr>
<td>LRI_strategy/Multi Multilinguals in LRI (n = 62)</td>
<td>3.66 (1.80)</td>
<td>6.40 (3.17)</td>
<td>2.73 (3.06)</td>
<td>0.89*</td>
</tr>
<tr>
<td>LRI_strategy/RP+ Strong readers in LRI (n = 59)</td>
<td>4.66 (2.43)</td>
<td>7.14 (3.00)</td>
<td>2.47 (3.11)</td>
<td>0.80*</td>
</tr>
<tr>
<td>LRI_strategy/RP- Weak readers in LRI (n = 59)</td>
<td>3.64 (1.87)</td>
<td>6.08 (2.63)</td>
<td>2.44 (3.03)</td>
<td>0.80*</td>
</tr>
</tbody>
</table>

With respect to strategy use, the results in Table 5 reveal that students of all language backgrounds benefit from the intervention LRI, as in every subgroup there are significant learning gains from pre- to post-test (with comparable and for all groups significant intra-group effect sizes between 0.70 and 0.89).

Table 6. Learning gains in items on noticing performance in LRI (adapted from Dröse, 2019, * marks a significant learning gain within the subsample)

<table>
<thead>
<tr>
<th>Scores in scale noticing</th>
<th>Pre-test m (SD)</th>
<th>Post-test m (SD)</th>
<th>Learning gain Δ m (SD)</th>
<th>Effect size d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRI</td>
<td>(n = 118)</td>
<td>2.11 (0.82)</td>
<td>2.48 (0.74)</td>
<td>0.38 (0.94)</td>
</tr>
<tr>
<td>LRI_noticing/Mono Monoilinguals in LRI (n = 56)</td>
<td>1.99 (0.81)</td>
<td>2.50 (0.71)</td>
<td>0.51 (0.87)</td>
<td>0.59*</td>
</tr>
<tr>
<td>LRI_noticing/Multi Multilinguals in LRI (n = 62)</td>
<td>2.21 (0.82)</td>
<td>2.47 (0.77)</td>
<td>0.26 (1.00)</td>
<td>0.26*</td>
</tr>
<tr>
<td>LRI_noticing/RP+ Strong readers in LRI (n = 59)</td>
<td>2.13 (0.81)</td>
<td>2.50 (0.71)</td>
<td>0.37 (0.90)</td>
<td>0.42*</td>
</tr>
<tr>
<td>LRI_noticing/RP- Weak readers in LRI (n = 59)</td>
<td>2.08 (0.84)</td>
<td>2.47 (0.78)</td>
<td>0.38 (0.99)</td>
<td>0.38*</td>
</tr>
</tbody>
</table>

With respect to the syntactic language awareness aspect of noticing syntactic features, the results in Table 6 reveal that students of all language backgrounds significantly benefit from
the intervention LRI (with intra-group effect sizes between 0.26 and 0.59), but less than for strategy use (with intra-group effect sizes between 0.70 and 0.89). Monolingual students tend to profit more for their noticing than multilingual students (learning gain 0.51 > 0.26 with effect size d = 0.59 vs. d = 0.26), briefly written LRI_notice/Mono > LRI_notice/Multi. In contrast, weak readers seem to profit equally for their noticing as strong readers (learning gain 0.38 and 0.38) and the difference between the effect sizes is not as high as for multilingual background (d = 0.38 vs. d = 0.42).

Table 7. Learning gains in items on interpreting performance in LRI (adapted from Dröse, 2019)

<table>
<thead>
<tr>
<th>Scores in scale interpreting</th>
<th>Pre-test m (SD)</th>
<th>Post-test m (SD)</th>
<th>Learning gain Δm (SD)</th>
<th>Effect size d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRI complete group</td>
<td>3.81 (1.26)</td>
<td>3.97 (1.17)</td>
<td>0.16 (1.45)</td>
<td>0.11</td>
</tr>
<tr>
<td>LRI_interpret/Mono</td>
<td>3.98 (1.05)</td>
<td>4.17 (1.13)</td>
<td>0.12 (1.36)</td>
<td>0.09</td>
</tr>
<tr>
<td>LRI_interpret/Multi</td>
<td>3.65 (1.42)</td>
<td>3.84 (1.21)</td>
<td>0.19 (1.53)</td>
<td>0.13</td>
</tr>
<tr>
<td>LRI_interpret/RP+</td>
<td>4.03 (1.07)</td>
<td>4.31 (0.98)</td>
<td>0.28 (1.33)</td>
<td>0.21</td>
</tr>
<tr>
<td>LRI_interpret/RP−</td>
<td>3.58 (1.40)</td>
<td>3.63 (1.26)</td>
<td>0.04 (1.56)</td>
<td>0.03</td>
</tr>
</tbody>
</table>

With respect to the syntactic language awareness aspect of interpreting syntactic features (when noticing is supported), the results in Table 7 reveal that students of all language backgrounds tend to benefit slightly from the intervention LRI (but not in significant ways), less than for strategy use and noticing. Monolingual and multilingual students equally profit for their interpreting (Learning gain 0.12 and 0.19 with effect size d = 0.09 and d = 0.13). Strong readers tend to profit substantially more for their interpreting than weak readers (learning gain 0.28 > 0.04) and effect sizes differ more than for multilingual background (d = 0.21 vs. d = 0.03).

Although the group comparisons do not reveal differences which are significant on a 5% level, the tendencies are worth being further explored and discussed.

5. Discussion

Students’ difficulties with word problems are well identified with respect to several obstacles. Whereas effective intervention programs have been developed with respect to conceptual and habitual obstacles (Verschaffel et al., 2000), and sometimes strategic obstacles (e.g., Reusser, 1994; Schukajlow et al., 2015), language obstacles on the word, sentence and text level have been mainly discussed with respect to reductive approaches of simplifying texts.

This paper, in contrast, reports on a language-responsive intervention aiming at fostering students’ word problem comprehension for overcoming strategic and language obstacles by instructional approaches. The design principles in view are strategic scaffolding and the variation principle for enhancing students’ language awareness. By combining strategy training with enhancing students’ syntactic awareness, we can enhance the current state of research on interventions for word problems (Zheng et al., 2013; Xin & Jitendra, 1999).

Additionally, we provide a first tentative contribution to researching which teacher support is required for implementation (Pellegrini et al., 2018; Saxe et al., 2001).
5.1 Limitations

The results of the evaluation study must be interpreted with caution due to its methodological limitations: The sample sizes were only small and not equally distributed so that significances between the two interventions could not be reached and random effects might still be too strong. Although the design was cluster-randomized, the assignment of only 22 teachers (and their classes) risks to cause further biases (2 more teachers had dropped out after the start due to personal reasons). Due to the small sample of teachers, it was not possible to include a multi-level structure with clustering at the teacher level into the data analysis. Further studies should involve at least 30 teachers for each treatment condition to conduct a multi-level analysis, this would allow to account for class effects. The larger sample sizes would also guarantee the splits according to language groups can still reveal more stable results and reach the desired significant level of 5%.

Additionally, future studies should also involve a control group with word problem set for students and no-treatment for teachers as a variation of NTG. Concerning the measures for the subscales on noticing and interpreting language obstacles, they need to be expanded in future studies in order to increase their reliability. Future studies should also expand the focus of the tested variables, in particular to investigate the connection to other aspects of mathematics performance.

In spite of these methodological limitations, the evaluation study reveals promising first insights into differential efficacies of the language-responsive instructional approach.

5.2 Summary and discussion of results

With respect to RQ1, the hypothesis of group differences in learning gains could be supported as the interaction effect $F_{\text{time} \times \text{group}}$ was highly significant. Additionally, although the alternative information-problems-only intervention worked with the same set of word problems and also sensitized teachers for the language obstacles in view, the language-responsive intervention tended to be more effective with higher learning gains in word problem comprehension. This supports the findings of Förster and Souvignier (2015) drawn about general reading comprehension intervention and provides first hints of necessary teacher support (Saxe et al., 2001). However, it should further disentangled in the use of professional development and support by curriculum material which be both varied between LRI and IPI in our first intervention (Hill & Charalambous, 2012).

Often, intervention studies are not reported if group differences are not significant. We decided to report these differences in effect sizes of the two interventions even though the significance level of 5% is not reached for many subresults. By this, we follow recent requests in the journal Nature even if we do not think that statistical significances should not be “retired” (Amrhein, Greenland, & McShane, 2019). Additionally, significance is often not reached in field studies with regular teachers (see Pellegrini et al., 2018; Cheung & Slavin, 2016). In spite of these limitations (and with the critical comments in mind), the reported tendencies justify the optimism that a larger evaluation study could confirm the efficacy of the language-responsive intervention and hence that identifying obstacles is not enough (see title).

For disentangling the efficacy with respect to different language backgrounds and different components of word problem comprehension, differential explorations were conducted. With
respect to research question RQ2, the effects of the two interventions were compared for each of the four language groups, monolingual and multilingual students, strong and weak readers. Each group tended to profit more from the language-responsive intervention than from the information-problems-only intervention. However, the learning gains were highest for the multilingual students. We interpret these tendencies as first indications that identifying obstacles is not enough for everybody, especially for multilingual students. This specifies the results of Förster and Souvignier (2015) for the slightly different learning content of subject-specific reading strategies, and it fulfills the call for investigating differential effectiveness for students with different language background (Snow, 1991).

In order to disentangle the students’ starting points, research question RQ3 compared the differences in pre-test scores in each scale for different language groups. Highest differences have been identified between monolinguals and multilinguals in strategy use, and between strong and weak readers in strategy use and interpreting. This supports the assumption that findings from previous more general studies can be transferred to mathematic-specific reading strategies and the processes of noticing and interpreting (Bos et al., 2012, Haag et al. 2013). Additionally, the results resonate with findings from a previous qualitative study on students’ strategies and language awareness (Dröse 2019).

Table 8. Overview on all comparative tendencies of this study

<table>
<thead>
<tr>
<th>(RQ1) Overall effectiveness of the two interventions</th>
<th>Findings</th>
<th>Comparison of effect sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1a) F_{time x group} with p &lt; 0.01 →</td>
<td>different learning gains between groups</td>
<td>( \eta^2 = 0.02 )</td>
</tr>
<tr>
<td>(1b) LRI &gt; IPI</td>
<td>d = 0.87 vs. d = 0.58 vs. d = 0.29</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(RQ2) Differential effects of the two interventions for different language groups</th>
<th>Findings</th>
<th>d = 0.82 vs. d = 0.42</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2a) LRI/Mono &gt; IPI/Mono</td>
<td>( \eta^2 = 0.02 )</td>
<td></td>
</tr>
<tr>
<td>(2b) LRI/Multi &gt; IPI/Multi</td>
<td>d = 0.91 vs. d = 0.67</td>
<td></td>
</tr>
<tr>
<td>(2c) LRI/RP+ &gt; IPI/RP+</td>
<td>d = 0.90 vs. d = 0.56</td>
<td></td>
</tr>
<tr>
<td>(2d) LRI/RP- &gt; IPI/RP-</td>
<td>d = 0.90 vs. d = 0.62</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(RQ3) Differences in pre-test scores in each scale for different language groups</th>
<th>Findings</th>
<th>p &lt; 0.01**</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3a) Strategy/Mono = strategy/Multi</td>
<td>p = 0.01*</td>
<td></td>
</tr>
<tr>
<td>(3b) Strategy/RP+ &gt; strategy/RP-</td>
<td>p = 0.02*</td>
<td></td>
</tr>
<tr>
<td>(3c) Noticing/Mono = noticing/Multi</td>
<td>p &lt; 0.01**</td>
<td></td>
</tr>
<tr>
<td>(3d) Noticing/RP+ &gt; noticing/RP-</td>
<td>p &lt; 0.01**</td>
<td></td>
</tr>
<tr>
<td>(3e) Interpreting/Mono &gt; interpreting/Multi</td>
<td>d = 0.80* and d = 0.80*</td>
<td></td>
</tr>
<tr>
<td>(3f) Interpreting/RP+ &gt; interpreting/RP-</td>
<td>d = 0.59* vs. d = 0.26*</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(RQ4) Differences in learning gains in each scale for different language groups within the language-responsive intervention</th>
<th>Findings</th>
<th>d = 0.70* and d = 0.89*</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4a) LRI_strategy/Mono &lt; LRI_strategy/Multi</td>
<td>d = 0.80* and d = 0.80*</td>
<td></td>
</tr>
<tr>
<td>(4b) LRI_strategy/RP’ = LRI_strategy/RP*</td>
<td>d = 0.59* vs. d = 0.26*</td>
<td></td>
</tr>
<tr>
<td>(4c) LRI_notice/Mono &gt; LRI_notice/Multi</td>
<td>d = 0.42* and d = 0.38*</td>
<td></td>
</tr>
<tr>
<td>(4d) LRI_notice/RP’ = LRI_notice/RP*</td>
<td>d = 0.09 and d = 0.13</td>
<td></td>
</tr>
<tr>
<td>(4e) LRI_interpret/Mono = LRI_interpret/Multi</td>
<td>d = 0.21 vs. d = 0.03</td>
<td></td>
</tr>
<tr>
<td>(4f) LRI_interpret/RP’ &gt; LRI_interpret/RP*</td>
<td>d = 0.09 and d = 0.13</td>
<td></td>
</tr>
</tbody>
</table>
Research question RQ4 concentrated on the language-responsive intervention and compared the learning gains of different language groups for each of the scales in an aptitude-treatment-interaction perspective (Snow, 1991). All language groups showed significant learning gains concerning strategy use and noticing syntactic features, with significant intra-group effect sizes. The differences between groups of different language background were not significant but still offer opportunities for generating hypotheses for further exploration. Monolingual and multilingual students have comparable learning gains in interpreting, but monolinguals tend to learn substantially more for their noticing than multilinguals, while multilingual learn more for their strategy use than monolinguals. In contrast, strong and weak readers have comparable learning gains in strategy use and noticing, but strong readers tend to profit more for their interpreting that weak readers.

In total, the significant difference to the no-treatment group and the differential tendencies confirm that strategy trainings are important for all students (as emphasized in earlier studies such as Reusser, 1994; Schukajlow et al., 2015). In contrast, the variation principle for enhancing students syntactic language awareness seems to be more beneficial for sensitizing strong readers’ interpreting than for weak readers. Even if the grammatical background knowledge for interpreting the structures was not provided in the intervention (as suggested by Schleppegrell, 2004), the variation principle was enough for those students to enhance their syntactic language awareness with respect to interpreting. This strengthens the optimism that a design principle which was introduced to mathematics education with respect to conceptual obstacles (Huang & Li, 2017) can also be activated for enhancing students’ syntactic language awareness. However, for multilingual students, it has to be combined with grammatical knowledge acquisition to enhance also their interpreting (Schleppegrell, 2004).

5.3 Practical consequences for the design of teaching materials and teacher professional development

These findings confirm the need for designing instructional interventions which support teachers and students to overcome typical challenges in word problems (Reusser, 1994; Verschaffel et al., 2010; Schukajlow et al., 2015).

However, the success of implementation will also depend on the teacher support (Saxe et al., 2001, Förster & Souvignier, 2015): Identifying the challenges alone (in research and in teacher preparation) is not enough, the discussion of design principles (in teacher professional development) and their integration in carefully designed teaching materials (in design research studies) seem to be crucial for linking research and practice more tightly (as requested by Heid et al., 2006, and many others). This result has substantial consequences for the need for more focused professional development on language-responsive interventions (as also requested by Erath et al., 2021).

Although the differential results (on language backgrounds influencing students’ strategies and language awareness and aptitude treatment interactions) might not yet be sufficiently stable due to limited reliability of the sub scales, the results already indicate that the intervention might be profitable for all students in a mainstream classroom with its usual language diversity: All student groups have considerable learning gains, even if they start with different aptitudes. However, the identified differences might also indicate that a differentiated look can strengthen
the effectiveness of instruction by putting different emphasis on strategy use and language awareness. Summing up, the intervention seems to have proven implementability, best with high teacher support and for all students. However, differentiated instruction for students with different language background might be crucial for leveraging each students’ competences for dealing with word problems.

Acknowledgments. This study was conducted within the MuM-research group (Mathematics learning under conditions of language diversity) at TU Dortmund University. It was partly funded by the German Ministry for Research and Education (Grant No. 03VP02270), partly by a grant by the German Telekom Foundation, both to S. Prediger. We thank our colleague Henrike Weinert for her statistical support.

6. References


