




Eye Movements During Mathematical Word Problem Solving—Global Measures and Individual Differences

Anselm R. Strohmaier  · Matthias C. Lehner · Jana T. Beitlich · Kristina M. Reiss

Received: 30 August 2018 / Accepted: 9 July 2019 / Published online: 24 July 2019
© GDM 2019

Abstract In mathematical word problem solving, reading and mathematics interact. Previous research used the method of eye tracking to analyze reading processes but focused on specific elements in prototype word problems. This makes it difficult to compare the role of reading in longer, more complex word problems and between individuals. We used global measures of eye movements that refer to the word problem as a whole, similar to methods used in research on eye movements during reading. Global measures allow comparisons of reading processes of word problems of different structure. To test if these global measures are related to cognitive processes during word problem solving, we analyzed the relation between eye movements and the perceived difficulty of a task and its solution rate. We conducted two experiments with adults and undergraduate students ($N=17$ and $N=42$), solving challenging mathematical word problems from PISA. Experiment 1 showed that more difficult items were read with shorter fixations, more saccades, more regressions, and slower, with correlations ranging from $r=0.70$ to $r=0.86$. Multilevel modelling in experiment 2 revealed that for the number of saccades and the proportion of regressions, the relationship was stronger for low-performing students, with performance explaining up to 37% of the variance between students. These two measures are primarily associated with building a problem model. We discuss how this approach enables the use of eye tracking in complex mathematical word problem solving and contributes to our understanding of the role of reading in mathematics.

Keywords Word problem · Reading · Eye tracking · Problem model · Performance · PISA

A. R. Strohmaier (✉) · M. C. Lehner · J. T. Beitlich · K. M. Reiss
Heinz Nixdorf Chair of Mathematics Education, TUM School of Education, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany
E-Mail: anselm.strohmaier@tum.de

Blickbewegungen beim Lösen mathematischer Textaufgaben – globale Maße und individuelle Unterschiede

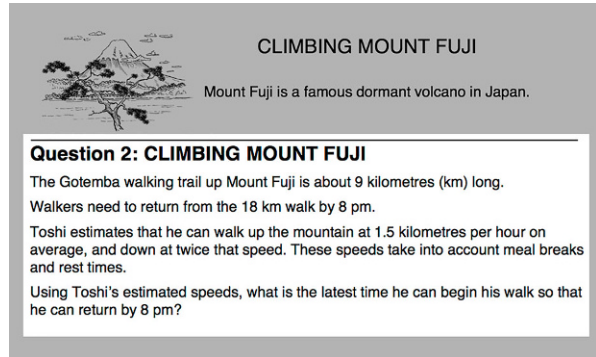
Zusammenfassung Beim Lösen mathematischer Textaufgaben interagieren Lesen und Mathematik. Bisherige Forschung verwendet die Methode des Eyetracking, um Leseprozesse zu analysieren, aber fokussiert auf spezifische, lokale Elemente in prototypischen Textaufgaben. Das macht es schwierig, die Bedeutung des Lesens zwischen längeren und komplexeren Textaufgaben sowie zwischen Personen zu vergleichen. Wir verwendeten globale Maße von Blickbewegungen, die sich auf die Textaufgabe als Ganzes beziehen. Diese werden etwa in der Forschung zu Blickbewegungen beim Lesen häufiger verwendet. Globale Maße ermöglichen einen Vergleich der Leseprozesse von Textaufgaben unterschiedlicher Struktur. Um zu prüfen ob diese Maße mit kognitiven Prozessen beim Lösen mathematischer Textaufgaben zusammenhängen, analysieren wir den Zusammenhang zwischen Blickbewegungen und der wahrgenommenen Schwierigkeit einer Aufgabe sowie ihrer Lösungsrate. In zwei Experimenten lösten Erwachsene und Studierende ($N=17$ und $N=42$) herausfordernde Textaufgaben aus PISA. Experiment 1 zeigte, dass schwierigere Items mit kürzeren Fixationen, mehr Sakkaden, mehr Regressionen und langsamer gelesen wurden, mit Korrelationen im Bereich zwischen $r=0,70$ und $r=0,86$. Mehrebenenanalysen in Experiment 2 zeigten, dass für die Anzahl der Sakkaden und den Anteil der Regressionen der Zusammenhang für schwächere Studierende anwächst, wobei Leistung bis zu 37% der Varianz zwischen den Studierenden aufklärt. Diese beiden Maße werden hauptsächlich mit dem Aufbau eines Problemmodells in Verbindung gebracht. Wir diskutieren wie dieser Ansatz die Nutzung von Eyetracking in komplexen Textaufgaben ermöglicht und zum Verständnis der Bedeutung von Lesen in Mathematik beiträgt.

Schlüsselwörter Textaufgabe · Lesen · Eye-Tracking · Modellieren · Leistung · PISA

1 Introduction

Mathematical word problems (WPs) play a prominent role in today's mathematics education (e.g., Boonen et al. 2016; Daroczy et al. 2015; Verschaffel et al. 2007). For example, WPs are used to assess mathematical abilities in large-scale studies like the *Programme for International Student Assessment* (PISA), as they allow the assessment items to be set within a real-world context and involve the application of mathematical concepts and skills (OECD 2013a; see Fig. 1). At the same time, students often struggle to understand WPs correctly (Cummins et al. 1988; Daroczy et al. 2015; Lewis and Mayer 1987; Nesher and Teubal 1975; Verschaffel et al. 2000). WPs are considered mathematical tasks in which relevant information is presented as text rather than in mathematical notation (Boonen et al. 2016; Daroczy et al. 2015; Verschaffel et al. 2000). Hence, both mathematical and reading abilities contribute to successful WP solving (for a review, see Daroczy et al. 2015). Specifically, reading

Fig. 1 Translated example stimulus that was used for both experiments (ID 5). Eye movements on the grayed-out area were omitted (the graying is for illustration and was not part of the stimulus). Adapted from PISA 2012 Released Mathematics Items (p. 20), by OECD (2013b). Copyright (2013) by the OECD. Used under CC BY-NC-SA 3.0 IGO



CLIMBING MOUNT FUJI

Mount Fuji is a famous dormant volcano in Japan.

Question 2: CLIMBING MOUNT FUJI

The Gotemba walking trail up Mount Fuji is about 9 kilometres (km) long. Walkers need to return from the 18 km walk by 8 pm.

Toshi estimates that he can walk up the mountain at 1.5 kilometres per hour on average, and down at twice that speed. These speeds take into account meal breaks and rest times.

Using Toshi's estimated speeds, what is the latest time he can begin his walk so that he can return by 8 pm?

is considered an immanent part of WP solving processes (e.g., Kintsch and Greeno 1985; Leiss et al. 2019).

The method of eye tracking has been used successfully to analyze the process of WP solving and the specific role of reading (De Corte et al. 1990; Hegarty et al. 1992; van der Schoot et al. 2009). However, this research commonly focuses on purely textual, specific prototype WPs, for example compare problems (Verschaffel et al. 1992) and on specific elements of the WPs such as keywords, numerical information or relational terms (Daroczy et al. 2015; De Corte et al. 1990; Hegarty et al. 1995). Thus, eye tracking has not been used to investigate how reading comprehension affects the process of solving more complex WPs and modelling tasks. Such WPs are often less structured and include context information, pictures and other representations.

In this study, we use eye tracking measures during WP solving that do not refer to specific elements of the WP. This way, we are able to compare reading processes between different items of different structure, length or layout, and between individuals. This comparison provides one of the key challenges in contemporary research on WP solving (Daroczy et al. 2015). For regular texts, models of reading explain how eye movements are associated with the difficulty of the text, characteristics of the reader and the cognitive processes underlying the reading process (Rayner et al. 2012). Building on Kintsch and Greeno's (1985) process model of WP solving, we argue which eye movements should be associated with the process of WP solving and WP difficulty. Moreover, we investigate whether this association varies with individual performance. We thereby take into account characteristics of both the WP and the WP solver and consider their interaction. This fills a research gap since it allows to use eye tracking for analyzing the role of reading in the process of WP solving and thereby compare individual reading patterns not only in prototype WPs, but in complex WPs as used in studies like PISA. Ultimately, the approach of integrating models from reading research in mathematics could offer new possibilities to analyze mathematical reading in a variety of contexts.

1.1 The Role of Reading Comprehension in Mathematical Word Problem Solving

Successful WP solving consists of at least three distinct processes (Nesher and Teubal 1975): (1) understanding of the linguistic structure of the WP, (2) understanding and constructing the relation between text and the arithmetic task, and (3) solving the arithmetic task. These steps can be followed consecutively in a linear, *direct translation strategy* (Hegarty et al. 1995). This might be efficient for some WPs (Pape 2004) but is more prone to inconsistency and linguistic complexity (de Koning et al. 2017; Hegarty et al. 1992, 1995; Lewis and Mayer 1987). Accordingly, WP solving is usually not a linear translation process (Boonen et al. 2013; Daroczy et al. 2015). Rather, successful students repeat and integrate the three steps in a *problem model strategy* (Hegarty et al. 1992; Leiss et al. 2019).

Kintsch and Greeno (1985) proposed a process model that describes how reading comprehension and problem solving interact in building a *problem model* during WP solving. According to the process model, the verbal input is transformed into a conceptual representation of its meaning and then organized to form the *text base*. From the text base, the problem model (or *situation model* in other research) is constructed by excluding irrelevant information from the text base and by inferring additional information. This process has shown to be a key predictor of successful WP solving, since the problem model contains the problem-relevant information in a form suitable for solution (Leiss et al. 2019). During these activities, reading comprehension and problem solving interact and do usually not occur step-by-step. Furthermore, bottom-up processes (retrieving textual information) and top-down processes (integrating cognitive resources) interact during WP solving (Leiss et al. 2019). Notably, this combination of bottom-up and top-down processes is also typical for text comprehension processes (e.g., Goldmann and Rakestraw 2000). Kintsch and Greeno's (1985) model has since been refined and modified by several researchers, but the basic processes remained fairly similar in most models of WP solving (e.g., Cummins et al. 1988; Nathan et al. 1992; Reusser 1990).

1.2 Reading Abilities and Mathematical Word Problem Solving

The process model by Kintsch and Greeno (1985) illustrates that the interaction of reading abilities and mathematical abilities is manifold. In their review, Daroczy et al. (2015) conclude that reading and mathematics cannot be considered isolated factors in WP solving and that therefore, WP solving skills are not merely the sum of reading comprehension for decoding and mathematical skills for calculating. The interaction of reading and mathematical processes forms a distinct factor that influences the difficulty of WPs (Abedi 2006; Boonen et al. 2016, 2013), and the process of mathematical modelling in particular (Leiss et al. 2010, 2019). This also means that students can successfully solve arithmetic tasks and have good reading comprehension skills but struggle when solving WPs (Daroczy et al. 2015).

Consequently, it is not possible to isolate processes of reading and mathematical thinking in WP solving. Rather, their interaction is characteristic for this kind of task (Daroczy et al. 2015). This also has implications for the assessment of WP

solving: Following the Kintsch and Greeno (1985) model, reading abilities cannot be considered a simple moderator or a possible linear control variable. Rather, the process of modelling is immanently influenced by reading (Leiss et al. 2019). The question if a particular student is a successful WP solver because of mathematical abilities or because of reading abilities is thus arguably not particularly meaningful, because their interaction is the key factor in WP solving. Similarly, controlling for reading abilities does arguably not provide researchers with the possibility to isolate WP solving abilities or the process of building a problem model, but rather obscures the observation (Daroczy et al. 2015). Accordingly, reading abilities are not merely a control variable, but distinctly influence WP solving (Leiss et al. 2019).

1.3 Individual Differences in Reading Mathematical Word Problems

Cognitive resources and prior knowledge are essential for WP solving, particularly because of the importance of top-down cognitive processes during reading and problem solving (e.g., Kintsch and Greeno 1985; Leiss et al. 2010, 2019). Verschaffel et al. (2000) describe several individual prerequisites for successful WP solving. They name a well-organized and flexibly accessible knowledge base, including knowledge about mathematics, linguistics, and the real world. Moreover, heuristic strategies are important for creating a meaningful problem model. Finally, they mention metacognitive strategies for self-regulation as well as affective components as important factors in WP solving. Moreover, both reading and mathematical thinking rely heavily on working memory and can interfere when one of the two processes causes a cognitive overload, which is especially likely in WP solving (Andersson 2007; Daroczy et al. 2015; Kintsch and Greeno 1985; Lee et al. 2009). Thus, differences in knowledge, knowledge structure and working memory affect WP solving (Riley et al. 1983). The variety of these prerequisites illustrates that individual differences in WP solving cannot be simply explained by mathematical skills, but that a broad range of individual characteristics influence the process of WP solving.

Prior research has shown that both reading comprehension and mathematical skills are associated with predominant WP solving strategies. Students who achieve lower in these domains focus on numbers and use a direct translation strategy more often compared to high-performing students, who are more likely to focus on words and use a problem model strategy (De Corte et al. 1990; Hegarty et al. 1995; Pape 2004). This indicates that the strategy that students use and their approach to solve WPs is often reflected in the way the WP is read (Hegarty et al. 1995).

From a theoretical standpoint, we thus assume that the processes of building a text base and a problem model challenge students differently and that these differences are reflected in the process and strategic approach of WP solving. High-performing students that are able to use a problem model strategy might tackle this multitude of challenges even in difficult WPs. But for students that rely on a direct translation strategy, building a text base and a problem model could be a major problem in solving WPs of increasing difficulty.

1.4 Eye Tracking and Mathematical Word Problem Solving

Reading is the ability to extract visual information and comprehend the meaning of texts (Rayner et al. 2012). Accordingly, key activities during reading are performed by the eyes. The method of eye tracking has been used to investigate the process of reading for over 130 years (see Rayner 1998; Rayner et al. 2012, for an overview). Consequently, the method has also been used to analyze the processes of learning (see Lai et al. 2013, for an overview), and specifically of WP solving in mathematics and science. A variety of studies illustrate that the semantic structure of WPs influences eye movement patterns during reading (e.g., De Corte et al. 1990; Hegarty et al. 1992, 1995; van der Schoot et al. 2009; Verschaffel et al. 1992).

Even though there are other kinds of eye movements, characterizing reading through *fixations*, *saccades*, and *regressions* is the basic tool for describing eye movements in reading (Rayner et al. 2012). During reading, the eyes do not move continuously, but make many short, rapid movements. Fixations last about 100 to 500 ms (Rayner 1998) during which time the eyes hardly move. At the end of a fixation, the eyes make a rapid, straight movement to the next fixated point. This movement is called a saccade. Perception is almost completely suppressed during saccades (Matin 1974). During fixations, information from a certain area around the fixated point can be processed. Since this area usually only covers a few characters, the eyes must gather information from a text step by step. Saccades that are directed opposite to the regular reading direction are called *regressive saccades* or *regressions* (Rayner et al. 2012). Typically, the eyes move about eight characters per saccade during regular reading. About 10 to 15% of all saccades are regressions (Rayner 1998).

Key results from research on eye movements during WP solving are that semantically more complex WPs are read slower when the mathematical content is controlled (De Corte et al. 1990; Hegarty et al. 1992). Depending on the semantic structure of WPs, specific elements like numbers, keywords, relational terms, and context information are fixated more often and longer (De Corte et al. 1990; Hegarty et al. 1992; van der Schoot et al. 2009). Semantically inconsistent WPs require more rereadings of parts of the problem that have already been read at least once (Hegarty et al. 1992; Verschaffel et al. 1992).

At the same time, studies on eye movements during WP solving support the assumptions that individual differences exist between students of differing skills, confirming that low-performing students focus more on numbers, reread problems more often and make use of a direct translation strategy more often (De Corte et al. 1990; Hegarty et al. 1992, 1995; van der Schoot et al. 2009).

While linguistic factors influencing mathematical WP solving have been investigated for various forms of WPs, studies using eye tracking methodology mostly focused on one-step and two-step addition, subtraction, multiplication, or compare problems (Daroczy et al. 2015). These prototype WPs offer a detailed perspective on very specific elements of the WP, such as relational terms, numerical information or semantic structures within the WP. According to the *eye-mind assumption* (Just and Carpenter 1980), the location of a fixation is closely related to the locus of attention. Hence, studies investigating these WPs usually use temporal and counting measures

of eye movements such as fixation times on these specific elements or the number of times they are fixated or refixated (Lai et al. 2013). We refer to measures that refer to specific elements of the WP as *local measures*. Typical local measures are the fixation time on numbers or the number of saccades between text and picture.

However, WPs used in textbooks, modelling tasks, or large-scale studies like PISA are often more complex and do not follow a prototype structure (Fig. 1). They use more text, include more context information or representations and often use a less formalized language. This complexity influences the process of WP solving (Daroczy et al. 2015; Jiménez and Verschaffel 2014; Leiss et al. 2010; Vicente et al. 2007). In more complex WPs, students usually need to identify relevant and irrelevant information, keep more information in the working memory and often need to carry out more complex calculations (Andersson 2007; Daroczy et al. 2015; Lee et al. 2009).

These processes cannot as easily be observed through local measures of eye movements. For example, items used in PISA vary regarding the mathematical content, text length and layout. Measures that refer to specific elements can therefore not be compared between different items. For example, fixation counts on keywords require WPs to be of similar length and structure in order to compare one problem to another.

1.5 Global Measures of Eye Movements

We refer to *global measures* of eye movements as parameters that do not depend on specific elements, areas of interest, or text length, but refer to the text as a whole. We use the global measures *mean fixation duration*, *number of saccades per word*, *proportion of regressions* and *reading speed*. The mean fixation duration is given by the average duration of all fixations on the WP. The total number of saccades is standardized by the number of words in the WP. The proportion of regressions is the number of regressions per overall saccades during reading. Reading speed is given by the number of words read per minute (for a technical description of the measures, see 2.1.8). Even though these measures are typically correlated with each other, they are assumed to relate to distinct cognitive processes during reading (Rayner et al. 2012).

Since global measures do not refer to specific elements but are standardized, it seems far less problematic to compare WPs of different length and structure (Rayner et al. 2012). This allows comparisons of reading processes between complex WPs. It has to be kept in mind that even though the measures are standardized, aspects like the average word length or the number of lines can vary between texts. Accordingly, there is an amount of variance remaining in global measures that cannot be attributed to the content (Rayner et al. 2012).

Global measures of eye movements have been used in research in mathematics education frequently, but only some studies have made use of them in the context of reading of mathematical texts. For example, reading speed has been used as a global measure of eye movements in mathematics education (De Corte et al. 1990). Inglis and Alcock (2012) found that expert mathematicians moved their gaze between lines (and arguments) more often when validating proofs compared to

novices. This number of between-line movements constitutes a global measure of eye movements, since it occurs irrespective of the specific content of the lines. It allowed Inglis and Alcock (2012) to compare reading processes of different proofs of varying content. Global measures of eye movements have further been used for analyzing reading processes of mathematical texts including illustrations, particularly in geometry (Epelboim and Suppes 2001). For example, by comparing mean fixation durations and saccades between text and illustrations, Lee and Wu (2018) analyzed the integration of geometric text and figures. Andrá et al. (2015) compared mean fixation durations between mathematical texts and formulae. Again, global measures allowed researches to compare elements of varying content and layout. In these studies, global measures of eye movements have been used to describe patterns of reading or reading strategies. In the present study, we use a model of reading for the interpretation of these measures.

Global measures of eye movements are typically used in research on reading of texts (Rayner et al. 2012). Here, the association between eye movements and cognitive processes underlying reading has been well described. In the current study, we transfer this association to WP solving. This enables us to use established global measures of eye tracking in a new way, which is the analyses of reading processes in complex WP solving based on the integration of process models of WP solving and reading.

1.6 Determinants of Global Measures of Eye Movements During Reading

The four global measures of eye movements show a specific variability among individuals and among different texts. In general, more difficult texts are read with longer mean fixation durations, more saccades, a higher proportion of regressions and at a slower reading speed (Rayner et al. 2006, 2012). At the same time, eye movements during reading depend on individual factors like reading abilities, reading intention, motivation or global strategy use (Radach and Kennedy 2013; Rayner et al. 2012). This illustrates that the cognitive effort required to process a text is reflected through global measures of eye movements during reading. It is unclear, however, if this association can be transferred to WPs, i.e., if the difficulty of a WP is associated with global measures of eye movements as well. If that is the case, these measures could be used to indicate cognitive processes during WP solving.

Research on reading provides elaborated models that explain this association (e.g., the E-Z Reader model; Reichle et al. 1998). The two key processes influencing eye movements during reading are *word recognition* and *text comprehension* (Clifton et al. 2016). We assume that similar processes can be identified in the process of WP solving. This will lead to our assumption that global parameters of eye movements during mathematical WP solving are associated with task difficulty and individual characteristics.

1.6.1 Word Recognition and Eye Movements

Simultaneous cognitive activities are involved in the process of word recognition. An important part of this relies on top-down processes, i.e. the reader makes use

of a prediction of the word to speed up reading. This prediction requires individual resources such as use of knowledge about surrounding words and inferring the context of the text, as well as working memory capacity. Hence, the “big three” variables that influence word recognition speed and success are *word frequency*, *word predictability* and *word length*: Words that are more frequent in the language of the text (e.g., “the”, “and”, or “have”) are fixated shorter and skipped more often than less common words, since they can be predicted more easily. The same is true for words that can be predicted by the context (e.g., “old” in “my cousin is one year old”) and for shorter words (Clifton et al. 2016; Rayner and Liversedge 2011). Words that are harder to recognize get reread more often and sometimes require more than one fixation. Primarily, these factors lead to an increase of mean fixation duration and a decrease in reading speed. Further, it increases the number of saccades and the proportion of regressions.

This paradigm can be transferred to WPs. In the process model of WP solving by Kintsch and Greeno (1985), the text base is first derived from the verbal input. This requires the transformation of the verbal input into a conceptual representation. This process is very similar to the process of word recognition. Accordingly, linguistic factors on the word-level, for example if the WP consists of longer, unfamiliar, or ambiguous words, influence WP difficulty (Hegarty et al. 1995; Daroczy et al. 2015) and should have an impact on the building of a text base. Because of these similarities, we assume that the process of building a text base is reflected in eye movements and related predominantly to mean fixation duration and reading speed.

1.6.2 Text Comprehension and Eye Movements

The second major determining factor of eye movements during reading is text comprehension. Higher order characteristics of a text such as its semantic or grammatical structure influence the reading process, for example in the moment an element of the text is not immediately understood by the reader (Clifton et al. 2016; Rayner et al. 2004). The eyes can further investigate the critical word (or critical region) until the problem has been resolved. This can lead to an increase in the number of saccades and proportion of regressions, longer fixation duration and a decrease in reading speed. Alternatively, the reader can execute a regression to reread previous segments of text—this leads to an increase in the number of saccades and proportion of regressions and again, a decrease in reading speed.

The more of these unexpected words, structures, or references occur, the more the reading is disrupted (Clifton et al. 2016). Text comprehension also depends on individual factors such as reading abilities, working memory span, motivation and individual reading intention (Radach and Kennedy 2004; Rayner et al. 2004). Compared to word recognition that mainly affects mean fixation duration, text comprehension should be reflected mainly through the number of saccades and proportion of regressions.

In WPs, the global structure of the WP, i.e. the semantic structure, the order of information, the placement of the question, or the presence of distractors influences the difficulty of a WP (Boonen et al. 2013; Daroczy et al. 2015). According to Kintsch and Greeno’s (1985) model, top-down comprehension processes are responsible for

organizing the text base, excluding irrelevant information, and inferring missing required information to build a problem model. Therefore, individual prerequisites like context knowledge, heuristic and metacognitive strategies, and affective components influence this process (Verschaffel et al. 2000). During reading of more complex WPs that are harder to comprehend, parameters of eye movements should be affected, as these processes are fairly similar to text comprehension processes during reading (Goldmann and Rakestraw 2000). Accordingly, the number of saccades and proportion of regressions should be related to processes of problem model building. It seems plausible that this should further contribute to a relation between global measures of eye movements and the process of WP solving.

In conclusion, global measures of eye movements during reading reflect underlying cognitive processes. Based on the process model by Kintsch and Greeno (1985) we assume that these measures similarly reflect cognitive processes during WP solving.

1.7 Summary and Hypotheses

We reviewed how WP solving requires not only the decoding of the numerical information embedded in the problem, but that reading comprehension and mathematical skills interact in successful WP solving. We discussed how global measures of eye movements are associated with cognitive processes during reading based on models of reading comprehension. Based on the process model by Kintsch and Greeno (1985), we transferred this framework to WP solving. We assume that global measures of eye movements can indicate the cognitive effort required to form a text base and to build a problem model. Since these measures are independent of the text length, content and layout, this would provide novel possibilities to analyze the process of WP solving.

We conclude by proposing the following hypotheses. First, we assume that when different WPs are compared, their difficulty should be related to global measures of eye movements during the solution process. Since similar cognitive processes occur during WP solving and reading of regular texts, we expect that global measures of eye movements are affected similarly. Thus, we expect that fixations during the solution of more difficult WPs are longer, that there are more saccades, a higher proportion of regressions, and a slower reading speed (H1).

Second, we expect that the relation between eye movements and task difficulty differs between individuals. The process of constructing a text base and building a problem model is increasingly important for more difficult WPs. Low-performing students tend to use a direct translation strategy and struggle with forming a text base and a problem model. Thus, their process of WP solving should be more impeded by an increase in problem difficulty. Accordingly, we expect that lower performing students are especially prone to show a pattern of a longer mean fixation duration, more saccades, a higher proportion of regressions, and a slower reading speed in WPs of increasing difficulty (H2).

2 General Method

2.1 Overview

The study consisted of two experiments. Experiment 1 was conducted as a pilot study with a smaller sample and a simplified design, only analyzing relations on the item level. Experiment 2 was planned and conducted to support data for more elaborate analyses, making use of multilevel modeling and thereby differentiating between the within-student level (L1) and the between-student level (L2). Overall, both experiments shared the same basic idea and general method.

2.2 Participants

We chose adults for our experiments who were proficient readers. In the present study, we did not assess prior abilities in mathematics or reading but were focusing on the relation between the solution process and WP performance. Assessing adults and not including an assessment of prior abilities gave us the possibility to include a larger number of participants. All participants gave written informed consent before participation. The study was conducted according to the *Ethical Principles of Psychologists and Code of Conduct* of the *American Psychological Association* from 2017. An ethics approval was not required by institutional guidelines or national regulations, in line with the guidelines of the *German Research Foundation*. A detailed description of the participants in the two experiments is given in Sect. 3.1 and 4.1, respectively.

2.3 Apparatus

Both experiments were conducted with a remote SMI RED 500 eye tracker. The device provides a sampling rate of 500 Hz. No head rest or other fixation was used. Participants sat in front of a 22-inch monitor with a resolution of 1680 × 1056 px in a distance of about 70 centimeters. The manufacturer's software *SMI iView X* and *SMI Experiment Center* were used for stimuli presentation and recording. *SMI BeGaze* was used for event detection.

2.4 Stimuli

In both experiments, participants solved nine WPs. Each WP was presented on one page. The stimuli consisted of published mathematical test items from the PISA-study (OECD 2006, 2013b). We rearranged the layout of items slightly so that they would fit on the screen and in such a way that the textual elements were isolated to reduce the accuracy requirements (e.g., pictures, see Fig. 1). The items were chosen to cover different areas of content (quantity, change and relationship, uncertainty and data) and mathematical processes (formulating, employing, interpreting; OECD 2013a). Items varied in their length between 1 and 11 lines, and included an irrelevant illustration (4 items) or a relevant illustration (5 items). All items featured a mathematical question that is embedded in real-world contexts like

hiking (see Fig. 1), a bike tour, a car sale, etc. Thereby, we covered a broad range of item characteristics, while making sure that none of the characteristics was directly related to item difficulty and could lead to a systematic bias. We used PISA items for several reasons. First, they are based on the concept of mathematical literacy and therefore cover a broad range of mathematical competencies beyond descriptive knowledge. Second, they represent WPs and include a certain amount of text, so they deliver sufficient data on reading processes. Most importantly, the tasks require knowledge that all students should have achieved by the end of their compulsory education. The influence of varying higher education is thereby reduced. Since the items are created for 15-year-olds, we chose items of above-average difficulty compared to the items used in PISA (about 0.3 *SD* higher) to assure that they provide an appropriate challenge for adult participants. This resulted in a mean solution rate of 69.8% in the present study. All stimuli were presented in German.

2.5 Task Difficulty and Performance

Participants' answers were coded correct or incorrect according to the scoring guide for the items (OECD 2006, 2013b). Depending on the aim of a study, participants' scores can be aggregated in two ways. The proportion of individuals who succeed in solving a task, i.e. the solution rate, is commonly called the difficulty of the task (Eccles and Wigfield 1995). We will refer to that difficulty as the *empirical task difficulty* (ETD). To improve readability, we inverted the solution rate so that a higher ETD corresponds to a more difficult item (0 representing a solution rate of 100% and 1 representing a solution rate of 0%).

In contrast, averaging the solution rate for each student results in a measure of performance. Experiment 1 focused on characteristics of the items, while experiment 2 took into account individual differences. Thus, we used the solution rate of items (ETD) in experiment 1 and the solution rate of participants (performance) in experiment 2. As usual, a higher solution rate corresponds to a higher performance score.

The difficulty of a mathematical task has various determinants apart from linguistic factors (Daroczy et al. 2015). The mathematical content knowledge that must be activated to solve problems, such as the number and nature of definitions, facts, rules, algorithms, and procedures is arguably the core determinant. In addition, the way and the context in which mathematical information is presented can influence the difficulty of the task (OECD 2014). All these factors accumulate in successful or unsuccessful solution of the WP. Therefore, the ETD might only make up for small differences in the reading process of WPs. We assume that the *perceived task difficulty* (PTD) that students report after solving an item might be more strongly associated with the process of building a text base and a problem model and hence with eye movements during reading compared to the ETD.

The PTD accounts for the fact that the difficulty during the solving process, especially for WPs, is not perfectly reflected by the solution (Hegarty et al. 1992). Therefore, the individual perception of difficulty should account for individual differences through the process of solving better than ETD. PTD has not been clearly

defined in research, although it has been considered an important factor in models of motivation (Eccles et al. 1985; Rheinberg et al. 2003).

PTD has been described as a two-dimensional concept, e.g., by Eccles and Wigfield (1995). The authors distinguished between task difficulty and required effort. In contrast, Rellinger et al. (1995) only consider the first of these two dimensions to constitute PTD and assess it with one single item that asks participants to rate how hard an item is compared to other tasks. Rheinberg et al. (2003) use a one-item scale similar to Rellinger et al. (1995) for the rating of the PTD. All three studies agree on this one dimension, but it remains unclear how much individual beliefs should be considered for PTD. We follow Rellinger et al. (1995) and Rheinberg et al. (2003), considering PTD as unidimensional. This includes only how difficult the task is perceived, and not how much effort is required to solve it.

To assess PTD, we therefore used a simple scale similar to Rellinger et al. (1995) and Rheinberg et al. (2003). Participants were asked to rate one question after each WP (“In general, I found this task ...”; 6-point Likert scale, 1 = “easy” and 6 = “difficult”). In the current study, we did not assess students’ prior abilities regarding their mathematical or reading abilities. Assessing prior knowledge would have increased individual testing time considerably and was not the theoretical focus of our study.

2.6 Procedure

Participants were made familiar with the apparatus. After the calibration, a validation was conducted. For the design of the experiment, the deviation was not as important as usual in research on eye movements during reading: The quality of all events that we assessed was independent of the absolute position of the event, for example the overall mean fixation time included all fixations on the text, irrespective of the position of the fixation. Therefore, validation only had to confirm that the area of the text was detected correctly. A 5-point calibration was repeated until a deviation of less than 1.0° was reached. This calibration links the data points obtained by the eye tracker to the coordinates of the stimuli. If an accuracy of less than 1.0° was not accomplished after three tries, the smallest mean deviation was used, and the data was checked afterwards to make sure that the deviation did not affect correct text area determination.

The items were presented in a randomized order. When an answer to the WP was given, the experimenter immediately presented the next slide containing the PTD scale, asking for the perceived difficulty of the previous item. Participants answered the WPs and the PTD-scale orally; the answer was recorded by the experimenter. The form of the required answer depended on the item and could be multiple choice or open answers, according to the coding of the original items. Participants were not allowed to take notes. This was necessary to assure that participants’ eye movements could be recorded during the whole process of WP solving. The challenges to integrate notetaking is a common issue in eye tracking research in mathematics education (e.g., André et al. 2015) and has to be considered in the interpretation of the results. There was no time limit on the tasks. After the experiment, participants answered a questionnaire conducting the background information. Lastly, they were

given the possibility to report their experiences during the experiment, e.g., if they were familiar with the tasks, and were debriefed. The experiment took about 55%, with participants spending about 45 min at the eye tracker. The time that students spend effectively working on the items was approximately 15 min, ranging from 8 min to 27 min. The rest of the time was spent with the calibration, introduction, and the PTD scales.

2.7 Event Detection

Items from PISA commonly include several graphical elements other than the WP text itself. For the detection of eye movement events, only text was included in the analysis while pictures, tables and graphs were omitted from analysis. This decision was based on the models of WP solving (Kintsch and Greeno 1985) and reading (Reichle et al. 1998) that both refer to verbal input for the building of a cognitive representation. This theoretical framework can therefore not necessarily be transferred to eye movements on representations. For example, it is unclear how the mean fixation duration on a graph should be interpreted regarding the building of a text base. To focus on mathematical aspects of the texts, we also limited event detection to the area of the mathematical problem and did not include introduction texts (e.g., “Mount Fuji is a famous dormant volcano in Japan” see Fig. 1).

Eye movement events were determined from the raw data in two steps. First, to detect fixations, we used the identification by dispersion (I-DT) algorithm proposed by Salvucci and Goldberg (2000). According to Blignaut (2009), a minimum fixation time of 100 ms and a maximum deviation threshold of approximately 2.5° were used.

During regular reading, fixations typically last for 100 to 500 ms (Rayner 1998). Thus, fixations longer than 500 ms were omitted. This accounts for the fact that participants’ eyes might rest on a point within the AOI for several seconds while he or she is thinking about the task but is not processing the text on this specific point. We tested afterwards if this procedure affected any results, which was not the case.

In the second step, saccades and regressions were detected. In general, dispersion algorithms detect only fixations (Holmqvist et al. 2011). Saccades are then calculated as eye-movements between fixations (SMI 2011). Since we were only interested in saccades during reading, we regarded saccades as the movement between two consecutive fixations that were both within the text. Typically, gaze positions are characterized with coordinates in pixels, with the top left corner having the coordinates (0,0). Hence, every saccade can be described by its horizontal (x-axis) and vertical (y-axis) extent. A regression is by definition directed opposite to the regular reading direction. Because of the amount of text, all items contained several lines. Detecting regressions in multi-line texts through an algorithm is not trivial since the direction of the horizontal movement is not a sufficient criterion. Moving to the start of the subsequent line (a return sweep, Rayner et al. 2012) results in a backward-directed saccade, whereas jumping to the end of a previous line (that has already been read) might result in a forward-directed regression. Therefore, every saccade was considered a regression if it was either 1) directed backward and less than 5 px downward (5 px was the interline spacing in the font size used for stimuli, i.e. the

distance between lines) or 2) directed forward but exceeded 20px upward (20px was the type size in the font size used for stimuli, i.e. the height of one line).

2.8 Eye Movement Parameters During Reading

To describe eye movements during reading, Rayner et al. (2012) uses four measures; the mean fixation duration, mean saccade length, proportion of regressions and reading speed in words per minute. Due to the specific requirements of our design, we adopted two of the measures but had to alter the other two.

We measured the mean fixation duration (MF) equivalent to studies on eye movements during silent reading. MF is the average fixation duration of all fixations on any part of the text, but only includes fixations shorter than or equal to 500 ms.

In a multi-line text, the measure of mean saccade length is not meaningful. Usually, this measure refers to the length of saccades in the regular forward reading process of a single line. In multi-line text, the reader might jump to a different part of the text, but the length of a saccade is not meaningful in this moment, since it could be short but aim for a semantically distant part of the text, or it could be very long but just aim for the beginning of the next line. We therefore used the total number of saccades, which is related to saccade length during regular reading, but not as prone to outliers. The number of saccades is standardized by the number of words of the item (SW).

Reading speed is usually given in words per minute (WPM). This is simply the number of words read in the allotted time. When working on WPs, this simple formula becomes problematic since reading might be interrupted by other activities, such as checking tables or thinking about the problem. As mentioned above, we accounted for this by only considering fixations on the text that did not last longer than 500 ms. Thus, we varied the WPM-measure slightly by dividing the number of words of the item by the total fixation time in minutes (WPFM).

Finally, the proportion of regressions in all saccades is given as a percent (PR).

It must be stressed that the task in our experiment differs to that usually used in experiments on eye movements during reading in one important way: While typically participants only have to decode and understand information given in a text, in our experiment participants also had to use this information to try to solve the mathematical content of the task. This must be accounted for when interpreting the

Table 1 Correlations between reading parameters

	MF (ms)	SW	PR (%)	WPFM
MF (ms)	–	0.76*	0.81**	–0.62
SW	0.65	–	0.76*	–0.90***
PR (%)	0.87**	0.89***	–	–0.55
WPFM	–0.60	–0.89***	–0.84**	–

Correlations for experiment 1 (n=9) are presented below the diagonal, and correlations for experiment 2 (n=9) are presented above the diagonal

MF mean fixation duration, SW saccades per word, PR proportion of regressions per saccade, WPFM words per fixation minute, ETD empirical task difficulty, PTD perceived task difficulty

*p < 0.05. **p < 0.01. ***p < 0.001

results, since parts of the text must be read several times before the information is decoded, evaluated and fully processed.

Typically, the four global measures are highly correlated for technical reasons. For example, a longer mean fixation time or more saccades decrease reading speed. For the example texts reported by Rayner et al. (2012), the four measures are correlated between $r=0.88$ and $r=0.97$ (calculated based on the reported descriptive data). In our dataset, correlations for both experiments are displayed in Table 1 and range from $r=0.60$ (MF and WPFM) to $r=0.90$ (SW and WPFM).

3 Experiment 1

In the first experiment, the general relationship between eye movement parameters and task difficulty on the item level was assessed. This was done in a simple correlational design with a smaller number of participants.

3.1 Method

Experiment 1 was conducted as described in the general method section. Participants were 17 adult members of the academic staff of a German university (11 female, 6 male) working in fields other than mathematics. All participants were native German speakers. Their mean age was 31.3 years ($SD=4.4$). Participants solved nine items.

3.2 Results

The rating of the PTD produced a mean value of 2.50 ($SD=1.25$). ETD ranged from 0 to 0.71 between items ($M=0.28$, $SD=0.21$) The descriptive reading and item parameters are displayed in Table 2. The mean fixation duration ranges from

Table 2 Descriptive reading and item parameters for experiment 1

Item	Reading Parameters				Item Parameters		
	MF (ms)	SW	PR (%)	WPFM	ETD	PTD	Words
ID 1	219	1.11	18.3	195	0.00	1.47	12
<i>ID 8</i>	226	1.38	23.8	173	0.06	1.65	43
<i>ID 6</i>	227	1.64	24.8	166	0.12	2.71	96
ID 2	228	0.86	22.7	262	0.18	1.76	17
<i>ID 5</i>	233	2.90	27.6	97	0.29	3.12	82
<i>ID 3</i>	225	1.31	21.3	194	0.35	2.88	45
<i>ID 4</i>	235	3.12	31.4	94	0.41	3.00	45
<i>ID 9</i>	227	2.33	27.2	99	0.41	2.65	23
<i>ID 7</i>	231	5.06	32.5	57	0.71	3.29	60

The items are displayed in order of their empirical difficulty, which is the inversed solution rate. Items in *italic* were used in both experiments

MF mean fixation duration, SW saccades per word, PR proportion of regressions per saccade, WPFM words per fixation minute, ETD empirical task difficulty, PTD perceived task difficulty

Table 3 Correlations between reading parameters and item parameters for experiment 1

	Reading Parameters				Item Parameters		Words
	MF (ms)	SW	PR (%)	WPFM	ETD	PTD	
ETD	0.61	0.86**	0.80**	-0.72*	1	0.81**	0.17
PTD	0.70*	0.75*	0.75*	-0.75*	0.81**	1	0.62

MF mean fixation duration, *SW* saccades per word, *PR* proportion of regressions per saccade, *WPFM* words per fixation minute, *ETD* empirical task difficulty, *PTD* perceived task difficulty

* $p < 0.05$. ** $p < 0.01$

219 to 235 milliseconds and is comparable to regular reading (Rayner et al. 2012). The proportion of regressions (PR) ranges from 18.3 to 32.5%, saccades per word (SW) range from 0.86 to 5.06, and the reading speed in words per fixation minute (WPFM) ranges from 57 to 262.

In Table 3, the correlations between the four measures of eye movements during reading and ETD and PTD are presented. ETD correlated significantly with SW ($r = 0.86$, $p < 0.01$), PR ($r = 0.80$, $p < 0.01$) and WPFM ($r = -0.72$, $p < 0.05$), whereas MF showed no significant correlation. PTD correlated significantly with all four reading parameters, with correlations ranging from $r = 0.70$ to $r = 0.75$ ($p < 0.05$). The directions of the association were as expected: The more difficult an item was (empirically or perceived), the longer the mean fixation time was, the more saccades were made per word, the bigger the proportion of regressions was and the slower the text was read. The four correlations with PTD and ETD did not differ pairwise ($Z < 0.55$, $p > 0.58$).

The two measures of difficulty correlated substantially ($r = 0.81$, $p < 0.01$). It should be noted that both ETD and PTD are related with the number of words in the task ($r = 0.17$, $p = 0.67$; $r = 0.62$, $p = 0.07$).

3.3 Discussion

The goal of experiment 1 was to investigate whether global eye movement measures are related to the difficulty of WPs. ETD was significantly related to the number of saccades per word, the proportion of regressions and reading speed. PTD was related to all four measures. Our assumption was that the processes described by the model of WP solving (Kintsch and Greeno 1985) influence eye movements similar to models of reading (Rayner et al. 2012). The first experiment supported this assumption. The two measures of difficulty were similarly related to eye movements and strongly related to each other. We therefore assume that rating the perceived difficulty of items on the individual level could be a good indicator for the underlying cognitive processes. This argument is limited by the fact that measures of difficulty were aggregated on the item level in experiment 1, thus it is not necessarily the case that they are related on the individual level.

Judging from this initial data, we were encouraged that the difficulty of a mathematical text might be connected to the eye movements during reading in a similar way as has been shown for the complexity of a regular text. Because of the promising results for PTD, we conducted a second experiment that allowed us to analyze the effect not only on the aggregated item level, but on the individual level.

Apart from that, we adjusted the design for the second study. One key limitation of the pilot study became obvious when the number of words was correlated with the two measures of difficulty. Both measures of item difficulty showed some association with the number of words in the task, even though this effect was not significant. The tendency was directed so that longer WPs were more difficult, both perceived and empirically. Even though our measures are independent of the text length from a technical perspective, a confounding of word count and text difficulty could mean that the difficulty of a task and eye movements during reading might correlate merely because of a shared connection to text length. Therefore, choosing the items for experiment 2, we minimized this correlation by changing two items that were relatively easy and consisted of fewer words compared to the other items, also dismissing one item that had been solved by all participants.

4 Experiment 2

4.1 Method

Experiment 2 followed the general method. Participants were 42 undergraduate students from a German University (15 female, 27 male) with an average age of 21.3 years ($SD=2.3$). They were students in programs other than mathematics (mostly mechanical engineering and education), 33 of them were in their first or second year at university. Apart from the reasons mentioned in 2.1.2, the decision to test students gave us the possibility to assess participants in an out-of-school context, voluntarily and in a low-stakes environment. This fit our efforts to analyze a thorough and deep WP solving process. Moreover, university students are assumed to have acquired basic skills to solve WPs, such as reading, basic computational skills and general knowledge about the problem contexts. All participants were native German speakers. Three item recordings were excluded from the data because of recording quality issues. Since two items were changed after experiment 1 (see 3.3), seven items were identical in experiment 2.

For our analysis, we chose a random coefficient regression model (Luke 2004) to account for the assumption that the association between task difficulty and eye movements during reading might vary between individuals. The model allows differentiation between effects on the within-students level (L1) and the between-students level (L2). The analysis was done in four steps (Aguinis et al. 2013). PTD was included as an L1 predictor, while students' performance (their overall score) was included as an L2 predictor. ETD was not included in this analysis, since using performance as an L2 predictor and the highly-confounded item score as L1 variable would influence the quality of the analyses. Both predictors were z -standardized for this analysis.

4.2 Results

Before the multilevel analyses were conducted, we analyzed the data in the same way as in experiment 1 to compare the results. Since two of the nine items were changed,

Table 4 Comparison of mean reading and item parameters between the two experiments

	Experiment 1 (<i>SD</i>)	Experiment 2 (<i>SD</i>)	<i>F</i> (1, 54)
MF (ms)	229 (20)	198 (31)	24.47***
SW	2.54 (1.92)	2.29 (1.46)	1.37
PR	26.9 (6.6)	30.1 (6.6)	12.34**
WPFM	147 (85)	187 (168)	2.90
PTD	2.76 (1.26)	2.53 (1.21)	2.48 ^a
ETD	0.24 (0.47)	0.31 (0.47)	0.08

Only the seven items that were used in both experiments are included in this comparison
MF mean fixation duration, *SW* saccades per word, *PR* proportion of regressions per saccade, *WPFM* words per fixation minute, *PTD* perceived task difficulty, *ETD* empirical task difficulty

^adf= 1,53

p*<0.01. *p*<0.001

Table 5 Correlations between reading parameters and item parameters for experiment 2

	Reading Parameters			Item Parameters			
	MF (ms)	SW	PR (%)	WPFM	ETD	PTD	Words
ETD	0.34	0.69*	0.31	-0.62	1	0.78*	0.13
PTD	0.68*	0.76*	0.43	-0.79*	0.78*	1	0.04

MF mean fixation duration, *SW* saccades per word, *PR* proportion of regressions per saccade, *WPFM* words per fixation minute, *ETD* empirical task difficulty, *PTD* perceived task difficulty

p*<0.05. *p*<0.01

we compared the mean values of all parameters considering only the seven items that had been run in both experiments. The results are given in Table 4. Participants in experiment 1 showed a significantly longer mean fixation duration, $F(1, 54)=24.47$, $p<0.001$. On the other hand, the proportion of regressions was significantly lower in experiment 1, $F(1, 54)=12.34$, $p<0.01$. The number of saccades per word and reading speed did not differ significantly, $F(1, 54)=1.37$, $p=0.25$; $F(1, 54)=2.90$, $p=0.09$. This was also true for the PTD rating and ETD, $F(1, 53)=2.48$, $p=0.44$; $F(1, 54)=0.08$, $p=0.78$.

The same correlations between the parameters of eye movements during reading were calculated and are displayed in Table 5. Slight differences compared to experiment 1 were observed: Now, correlations for ETD emerged smaller, only SW still correlating significantly ($r=0.59$, $p<0.05$). On the other hand, the correlation between the proportion of regressions and the perceived task difficulty was not replicated. Apart from that, the correlations were similar to experiment 1. The correlation between the number of words and the ETD and PTD could be decreased to $r=0.12$ and $r=0.04$ respectively. The direction of the correlations emerged as reported for experiment 1.

The results from the random coefficient regression models are displayed in Tables 6, 7, 8 and 9. ICCs indicate how the overall variance in reading parameters is distributed between L1 and L2. For MF, the between-students level (L2) accounted for 69% of the variance, which indicates strong individual differences. In contrast, L2 only accounts for 6% of the overall variance for SW, which indicates that most of the variance is within students. L2 accounts for 18% of the variance in the pro-

Table 6 Results of multilevel modeling analysis for mean fixation duration (MF) in ms

Level and Variable	Model			
	Null (Step 1)	Random Intercept and Fixed Slope (Step 2)	Random Intercept and Random Slope (Step 3)	Cross-Level Interaction (Step 4)
Level 1 (within)				
Intercept ($\hat{\gamma}_{00}$)	196.9** (4.2)	196.9** (4.2)	196.9** (4.2)	196.8** (4.2)
PTD ($\hat{\gamma}_{10}$)	–	4.5** (1.2)	4.4** (1.2)	4.3** (1.2)
Level 2 (between)				
Performance ($\hat{\gamma}_{01}$)	–	3.0 (4.2)	1.8 (4.5)	2.7 (4.3)
Cross-Level				
PTD Performance ($\hat{\gamma}_{11}$)	–	–	–	–1.2 (0.8)
Variance components				
Within-person (L1) variance ($\hat{\sigma}^2$)	313.9	297.3	288.2	286.9
Intercept (L2) variance ($\hat{\tau}_{00}$)	691.4	702.8	707.5	699.6
Slope (L2) variance ($\hat{\tau}_{11}$)	–	–	10.6	12.0
Intercept-slope (L2) covariance ($\hat{\tau}_{01}$)	–	–	48.4	–48.2
Additional Information				
ICC	0.69	–	–	–
Explained variance				
Within-person (L1)	–	0.05	0.08	0.09
Intercept (L2)	–	0.00	0.00	0.00
Slope (L2)	–	–	–	0.00

Values in parentheses are standard errors. *t*-statistics were computed as the ratio of each regression coefficient divided by its standard error

PTD perceived task difficulty

** $p < 0.01$

portion of regressions. This might explain why the correlation between PR and PTD was not found in our first analysis, as notable variance exists between individuals.

Including the two predictors in a random intercept and fixed slope model led to similar results for all four outcome measures. In every case, PTD predicted reading parameters significantly within students. Items that were perceived more difficult were read with longer fixations, more saccades, a higher proportion of regressions, and slower. In contrast, performance was not related to reading parameters.

In model 3, slopes were allowed to vary between individuals. The variance of the slopes was then used as a baseline measure for model 4, when performance was included as a predictor of the L1 effect of PTD and reading parameters. This cross-level interaction effect is displayed in Fig. 2. It was significant only for PR and SW, but not for MF and WPFM. For PR, this effect accounted for 37% of the variance in slopes between students. In contrast, none of the variance in slopes was explained with MF as outcome variable. For both PR and SW, a higher performance led to

Table 7 Results of multilevel modeling analysis for saccades per word (SW)

Level and Variable	Model			
	Null (Step 1)	Random Intercept and Fixed Slope (Step 2)	Random Intercept and Random Slope (Step 3)	Cross-Level Interaction (Step 4)
Level 1 (within)				
Intercept ($\hat{\gamma}_{00}$)	2.168** (0.087)	2.170** (0.090)	2.157** (0.100)	2.148** (0.098)
PTD ($\hat{\gamma}_{10}$)	–	0.770** (0.086)	0.728** (0.086)	0.722** (0.083)
Level 2 (between)				
Performance ($\hat{\gamma}_{01}$)	–	–0.045 (0.083)	–0.004 (0.091)	–0.049 (0.079)
Cross-Level				
PTD Performance ($\hat{\gamma}_{11}$)	–	–	–	–0.160* (0.070)
Variance components				
Within-person (L1) variance ($\hat{\sigma}^2$)	1.775	1.209	1.088	1.084
Intercept (L2) variance ($\hat{\tau}_{00}$)	0.116	0.201	0.210	0.206
Slope (L2) variance ($\hat{\tau}_{11}$)	–	–	0.131	0.117
Intercept-slope (L2) covariance ($\hat{\tau}_{01}$)	–	–	0.074	0.072
Additional Information				
ICC	0.06	–	–	–
Explained variance				
Within-person (L1)	–	0.32	0.39	0.39
Intercept (L2)	–	0.00	0.00	0.00
Slope (L2)	–	–	–	0.11

Values in parentheses are standard errors. *t*-statistics were computed as the ratio of each regression coefficient divided by its standard error

PTD perceived task difficulty

p* < 0.05. *p* < 0.01

a decrease in the regression slope, so the better students performed, the smaller was the relationship between PTD and the reading parameter.

Overall, the models accounted for none of the L2 variance for MF, SW and WPFM and for 19% of the L2 variance for PR. On L1, between 8 and 39% of the variance could be explained by including the two predictors and the cross-level interaction.

4.3 Discussion

In experiment 2, differences in the mean reading parameters occurred compared to experiment 1. Participants showed longer mean fixation durations and a slower reading speed, but also a smaller proportion of regressions. This seems to reflect a general difference in the reading pattern between participants. While the academic

Table 8 Results of multilevel modeling analysis for the proportion of regressions (PR) in percent

Level and Variable	Model			
	Null (Step 1)	Random Intercept and Fixed Slope (Step 2)	Random Intercept and Random Slope (Step 3)	Cross-Level Interaction (Step 4)
Level 1 (within)				
Intercept ($\hat{\gamma}_{00}$)	28.79** (0.58)	28.79** (0.54)	28.80** (0.53)	28.71** (0.54)
PTD ($\hat{\gamma}_{10}$)	–	1.70** (0.42)	1.62** (0.43)	1.58** (0.41)
Level 2 (between)				
Performance ($\hat{\gamma}_{01}$)	–	0.46 (0.53)	0.51 (0.54)	0.48 (0.53)
Cross-Level				
PTD Performance ($\hat{\gamma}_{11}$)	–	–	–	–0.78* (0.33)
Variance components				
Within-person (L1) variance ($\hat{\sigma}^2$)	41.24	39.30	37.94	37.81
Intercept (L2) variance ($\hat{\tau}_{00}$)	9.24	7.90	7.02	7.52
Slope (L2) variance ($\hat{\tau}_{11}$)	–	–	1.98	1.24
Intercept-slope (L2) covariance ($\hat{\tau}_{01}$)	–	–	0.04	0.17
Additional Information				
ICC	0.18	–	–	–
Explained variance				
Within-person (L1)	–	0.05	0.08	0.08
Intercept (L2)	–	0.15	0.24	0.19
Slope (L2)	–	–	–	0.37

Values in parentheses are standard errors. *t*-statistics were computed as the ratio of each regression coefficient divided by its standard error

PTD perceived task difficulty

* $p < 0.05$. ** $p < 0.01$

staff members in experiment 1 read slightly faster, but less steadily, the students in experiment 2 read more carefully and, consequently, had to backtrack less often. This finding indicates that there might be systematic differences in reading patterns between different populations, which should be kept in mind when generalizing the findings. In particular, it is very likely that reading parameters would differ between our adult sample and a sample of 15-year-olds for which the items were initially designed.

Experiment 2 confirmed the key finding from experiment 1 that three of our four parameters of eye movements during reading are associated with the PTD but failed to replicate the effect on the proportion of regressions. In fact, the correlation emerged notably smaller than in the first experiment. This might be an indicator that the number of words might indeed have been a confounding factor for this correlation in experiment 1. At the same time, the fact that the elimination of this confounding

Table 9 Results of multilevel modeling analysis for words per fixation minute (WPFM)

Level and Variable	Model			
	Null (Step 1)	Random Intercept and Fixed Slope (Step 2)	Random Intercept and Random Slope (Step 3)	Cross-Level Interaction (Step 4)
Level 1 (within)				
Intercept ($\hat{\gamma}_{00}$)	191.4** (12.6)	191.4** (12.9)	198.8** (17.7)	199.7** (13.4)
PTD ($\hat{\gamma}_{10}$)	–	–55.6** (8.8)	–43.7** (11.6)	–42.3** (8.4)
Level 2 (between)				
Performance ($\hat{\gamma}_{01}$)	–	–10.1 (10.2)	5.8 (8.1)	–8.2 (10.7)
Cross-Level				
PTD Performance ($\hat{\gamma}_{11}$)	–	–	–	16.4 (8.4)
Variance components				
Within-person (L1) variance ($\hat{\sigma}^2$)	22030.7	19239.9	17358.3	17334.8
Intercept (L2) variance ($\hat{\tau}_{00}$)	4455.1	4846.8	6148.4	5744.0
Slope (L2) variance ($\hat{\tau}_{11}$)	–	–	1207.2	1019.6
Intercept-slope (L2) covariance ($\hat{\tau}_{01}$)	–	–	–2723.8	–2419.5
Additional Information				
ICC	0.17	–	–	–
Explained variance				
Within-person (L1)	–	0.13	0.21	0.21
Intercept (L2)	–	0.00	0.00	0.00
Slope (L2)	–	–	–	0.15

Values in parentheses are standard errors. *t*-statistics were computed as the ratio of each regression coefficient divided by its standard error

PTD perceived task difficulty

***p* < 0.01

factor did not influence the other three correlations supports our assumption that they are related to the difficulty of the task.

For ETD, the correlation with the number of saccades per word was replicated, while reading speed and the regression ratio did not correlate significantly in experiment 2. The drop of the correlation between ETD and the proportion of regressions could support the assumption that the students in experiment 2 read more steadily and used regressions as a tool for WP comprehension less often. Still, this result might also be an artifact from the change of two of the nine items.

As expected, results from the multilevel analysis confirmed that the relationship between PTD and parameters of eye movements during reading persists on the within-student level. Explaining between 8 and 39% of the variance in these measures, the models indicate a strong relationship between how difficult a WP is perceived and how the eyes move during reading. Our hypotheses were partly confirmed for the between-student level: Performance did not have a main effect on

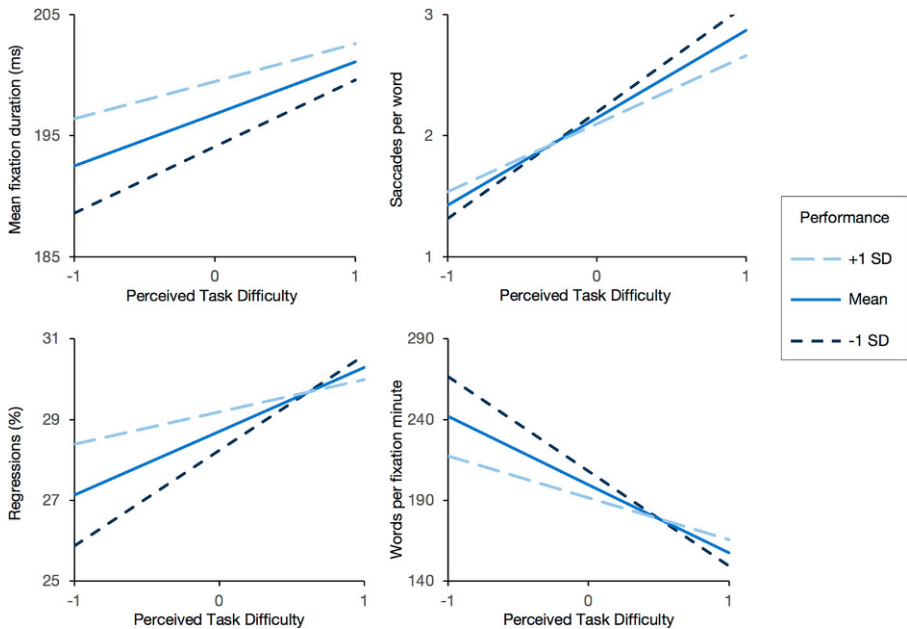


Fig. 2 Plots of moderating effects of L2 variable achievement on the relationship between L1 variable perceived task difficulty (PTD, z-standardized) and L1 variables mean fixation duration (in ms), saccades per word, proportion of regressions (in percent), and words per fixation minute

reading parameters, but a cross-level interaction effect was found for the proportion of regressions and SW. The association between PTD and these two parameters is significantly stronger for low-performing students than for high-performing students. That means that the perceived difficulty is related to the process of reading more strongly in low-performing students.

5 General Discussion

5.1 Key Findings

In the two experiments, we found support for our hypotheses. In H1, we expected that the process of WP solving is reflected in the four global measures of eye movements. Specifically, we hypothesized that the process of building a text base and a problem model should influence mean fixation duration, number of saccades, proportion of regressions and reading speed. In more difficult WPs, this should lead to a change in these measures. This was indeed the case and will be further discussed in Sect. 5.3 to 5.4.

In H2, we hypothesized two things: The process of WP solving depends on a variety of individual prerequisites that go beyond mathematical knowledge or skills. Thus, we expected that the relation between eye movements and task difficulty differs between participants. This was reflected in our data. Second, we expected that

is especially the lower achieving students that are affected by the requirements of building a text base and a problem model. Therefore, we expected that for students with lower overall scores in the WPs, the perceived task difficulty should be stronger related to the four global measures of eye movements. We this effect to be significant for two of the measures (SW and PR). Implications from these findings will be discussed in Sect. 5.5.

It has to be taken into account that all findings refer to our sample of adults. Our results indicate that the method described is a valuable tool in analyzing the process of WP solving, but it remains unclear to what extent the results can be generalized to younger students.

5.2 Parameters of Eye Movements During Reading of Mathematical Word Problems

Compared to the reference reading parameters from Rayner et al. (2012), the descriptive data offers a good first impression of the global measures of eye movements during the reading process of WPs compared to regular texts. The mean fixation duration is equivalent to that of a text of medium complexity. The other three measures are influenced when the reader re-reads the task, which is very common for a WP (Hegarty et al. 1992) but was probably more uncommon in the study reported by Rayner et al. (2012). Accordingly, the proportion of regressions is at the top end of what Rayner et al. (2012) observed during regular silent reading. Since regressions seem to be a key tool for assessing mathematical texts (Inglis and Alcock 2012), this appears reasonable. For the two measures saccades per word and reading speed, the numbers cannot be easily compared to the reference. Still, with a mean word length of 5.4 characters for the WPs, the saccades per word correspond to mean saccade lengths of about 1 to 6.5 characters, which also starts at the bottom end of what is common during regular reading. When comparing words per minute (WPM) and words per fixation minute (WPFM), WPFM would emerge higher since the total reading time is bigger than the sum of fixation times. Still, compared to readers of unspecific texts, participants in our study read slower. This might simply be caused by repeated go-throughs.

Similar to eye tracking research on the reading of other texts, the measures that we used were substantially correlated. Even though we argued that they build on distinct cognitive processes and can therefore not be used interchangeably, it seems problematic to interpret them in isolation. The relation between the parameters should always be considered in the ultimate interpretation of the data. Accordingly, even though we did separate analyses for all measures, our interpretation will always refer to their combination.

5.3 The Relation Between Word Problem Difficulty and Reading Parameters

The first goal of this study was to investigate whether the paradigm that the difficulty of a text is associated with parameters of eye movements during reading can be expanded to WPs. We found a similar relationship in two experiments which supports H1, though the details, limitations and applications must be discussed.

The results of the two experiments concur in the finding that in more difficult WPs, mean fixation duration is higher, there are more saccades and a higher proportion of regressions and the WP is read slower. In addition, experiment 2 confirms this effect on the within-student level for PTD on all four measures. The direction of the relationship is in line with findings about the relationship of eye movements and text difficulty during reading (Rayner et al. 2012). At the same time, the findings align with existing studies regarding eye movements during the reading of mathematical texts: Similar to prior studies, global measures of eye movements were linked to cognitive processes during reading (Andrá et al. 2015; De Corte et al. 1990; Epelboim and Suppes 2001; Hegarty et al. 1992, 1995; Inglis and Alcock 2012; Lee and Wu 2018; van der Schoot et al. 2009; Verschaffel et al. 1992). The results are also in line with Daroczy et al. (2015) since the relation of reading for the process of solving WPs was supported. Still, by integrating a model of reading and a model of WPs solving we believe that our approach offers a new way for the interpretation of eye movements in reading of mathematical texts, specifically for complex WPs.

5.4 Causality and Cognitive Processes

Our participants read tasks that they perceived more difficult with longer fixations, more saccades, a higher proportion of regressions, and more slowly. At first, this relation does not imply causality in either direction. It is possible that items that are read similarly to complex texts are harder to solve or at least perceived to be more difficult, because it is harder for the reader to extract and process the relevant information. Still, this seems unlikely. For example, a longer mean fixation duration should support word recognition, not hinder it. Furthermore, it seems unreasonable that a reader should struggle with a task because of a slow reading pace. Similar arguments can be found for the other parameters of eye movements.

On the other hand, it is more likely that both the mathematical content and the linguistic structure of the text influence reading in the same way that is explained by models of eye movements during reading: Here, the relationship between text difficulty and eye movements during reading is explained through the two key factors word recognition and text comprehension. As we briefly reviewed, both factors require more cognitive resources and make it harder to read efficiently when words are less common, and the linguistic structure is more complex. This increased effort is then reflected in the eye movements. We hypothesized that the activities described by the model of mathematical WP solving by Kintsch and Greeno (1985) offers a good explanation why a similar relation can be expected here.

On the one hand, the process of building a text base should be closely related to processes of word recognition, since the verbal input needs to be transformed and organized into a conceptual representation. More difficult mathematical tasks include longer, unfamiliar or ambiguous words, concepts, and numbers that are harder to process and thereby hinder word recognition. At the same time, word recognition is known to affect eye movements, mainly mean fixation duration and reading speed.

On the other hand, inferring needed information and excluding irrelevant information to form the problem model requires text comprehension. The more complex linguistic and mathematical structure, the order of information, or the presence of

distractors makes it harder to build a syntactic and mathematic representation and to comprehend more difficult mathematical tasks (Daroczy et al. 2015). These processes of text comprehension are mainly associated with the number of saccades and the proportion of regressions.

Even if our results support the assumption of a causal relation between WP difficulty and eye movements, our explanation could be further investigated. A more qualitative approach that focuses in detail on the events within the reading process that cause variation in global reading parameters might contribute to a better understanding of the details of the process.

5.5 Individual Differences

Several studies concur that WP solving differs between students of varying mathematical and reading comprehension skills (Hegarty et al. 1995; Leiss et al. 2010, 2019). Our results show that performance is associated with the relationship between eye movements during reading and task difficulty for the proportion of regressions and number of saccades, partly confirming H2. In our framework, we associated these two parameters with the process of building a problem model, which heavily relies on processes of text comprehension. This association is of great importance, because it indicates that linguistic factors contributing to WP difficulty have a bigger impact on low-performing students. It implies that for low-performing students, linguistic processing and comprehension is perceived as one key difficulty-generating factor of WP solving, while higher performing students might focus more on other aspects of the WP, such as the mathematical content. This is in line with findings that lower performing students make use of a direct translation strategy more often which is prone to semantic inconsistencies or complex semantic structures (Inglis and Alcock 2012; van der Schoot et al. 2009; Verschaffel et al. 1992). These students might then struggle with building a problem model in more difficult items when a direct translation strategy becomes more and more unfavorable.

In contrast, no interaction effect emerged for mean fixation duration and reading speed, which we associated with the process of building a text base. Even though our previous results show that there is a relation between these parameters and WP difficulty, it did not systematically depend on achievement in our sample. This might indicate that this process is an important factor in WP solving, but in a similar way for all students.

The analysis of the process of WP solving through eye movements could provide valuable insights when prior reading and mathematics abilities are taken into account. Leiss et al. (2019) found that reading abilities show an indirect effect on performance in complex WPs, mediated through strategies in building a problem model. In addition to the think-aloud method used by Leiss et al. (2019), our interpretation of eye movements could help clarify this indirect relation.

These findings have important implications for instructional practice. The challenges of WP solving seem to differ between students of varying performance, with linguistic aspects being more important in low-performing students. Partly, this might be caused by abilities and strategies that can be improved, for example by providing students with tools for acquiring a problem model strategy, metacognitive

reflection or reading comprehension. Eye tracking could serve to indicate the success of such efforts. Moreover, individual differences in eye movements could help identify factors influencing the process of WP solving that might not necessarily be visible in the WP solution.

5.6 Application

Before generalizing our findings, it has to be discussed to what extent our results might be specific for our population of adults. The general association between eye movement parameters and task difficulty is probably rather universal, but the fact that we found individual differences regarding its extent could mean that younger students or students that are not proficient readers could show a different relation (Rayner et al. 2012). Since the association increased in lower-performing adults, it is likely that students who typically perform slightly worse on the same WPs should also show a substantial relation between perceived task difficulty and global measures eye movements. However, it is also possible that perceived task difficulty is reported differently by younger students or that mathematical operations are of bigger importance in WP solving for them, which could reduce the relation.

We believe that these findings offer a variety of applications. Most importantly, using global measures of eye movements allows analyzes and comparisons of longer, more complex WPs that differ regarding their content, length and layout. This offers new ways of using the method of eye tracking in mathematics education research, particularly for analyzing mathematical texts. At the same time, using global measures comes with technical advantages. Since the absolute position of the gaze becomes less important compared to the relative movement and duration, the accuracy of the eye tracking device is less crucial compared to local measures. This could, for example, allow the use of less precise eye-tracking glasses in a more authentic environment. Moreover, the whole text is considered as AOI, which avoids the question of how AOIs are defined and how measures are compared between these AOIs. For example, the question if fixation time needs to be standardized by the size of the AOI becomes obsolete.

The use of our method can be expanded beyond the basic analyses reported in this paper. It can be used to characterize how linguistic factors influence the solution process and the difficulty of complex mathematical WPs, and to better understand who is affected by these components. By systematically manipulating mathematical or linguistic features of a WP, the assumption made by Nesher and Teubal (1975), Daroczy et al. (2015), and others that linguistic and mathematical features of a WP interact in composing difficulty could be further investigated. Similarly, a differentiation between items and between individuals as postulated by Daroczy et al. (2015) can be achieved with this approach, as it allows comparison of different types of WPs.

The method offers a tool for assessing the difficulty of a mathematical task non-invasively and on-line. To a certain degree, the question of how difficult a task is perceived to be by a person can be answered by using the eye movements during the solving process. This can be used in combination with task responses, self-reports, or think-aloud protocols. The benefit of eye tracking lies in the possibility

to observe the solution process directly and in detail. Compared to other methods of observation, this often provides additional information (e.g., Schindler and Lilienthal 2019). At the same time, the approach is not limited to mathematical WPs but is applicable to any kind of word problems. Like the concept of mathematical literacy, recent conceptions of abilities in many domains include real-world applications and therefore make use of word problems, specifically science education.

Furthermore, our approach could be transferred to mathematical texts other than WPs, for example school book texts. With the rapid development of the eye tracking method, innovative learning formats like games, e-books or collaborative learning platforms that contain mathematical texts could be an interesting field of research, especially since the method is already prominent in these fields. But also, the development of eye tracking glasses could offer possibilities to assess reading processes during the work with more traditional, paper-based material.

Regarding our results, the fact that individual differences exist in the relationship between eye movements during reading and WP difficulty indicates that this relationship could be used as a measure of how much a person is affected by linguistic factors when solving WPs. This opens a new perspective on comparing WP solving between different groups of students and between languages. When extended to more items, the interrelation with the mathematical content could be assessed. For example, in geometry WPs, reading comprehension might have a smaller influence on its perceived difficulty than in algebraic WPs. Furthermore, processes of WP solving in other subjects than mathematics could be compared to our findings.

We briefly discussed how we believe that the integration of the four measures of eye movements during reading can be used to characterize global reading strategies. Even though this assumption cannot be verified within this study, we propose a possible interpretation for the differences between experiment 1 and experiment 2. Academics applied a more unsteady reading strategy, relying on regressions to make up for missed information. This could be a sign of a problem model strategy. On the other hand, students might have read more thoroughly, which might indicate a direct translation strategy (Hegarty et al. 1992). Future studies could investigate this population differences, possibly identifying systematic influences (e.g., age). At this point, our data cannot prove this assumption. Refined, the method could be used to search for successful reading strategies for WPs by analyzing inferences between the different measures of eye movements and relate them to solutions, other eye tracking data or think-aloud protocols.

5.7 Limitations

Our approach contrasts with former research on the reading of mathematical texts. It adopted methods and knowledge from research on global measures of eye movements during reading, which in our view offers powerful possibilities, but also some disadvantages.

In opposition to most existing research that made use of eye tracking, we did not differentiate between specific elements of the WP. We considered the problem as one text and used global measures of eye movements on this whole text. Thereby, we lose the possibility to analyze specific mathematical strategies like the relation of text

and numbers or the role of relational terms. But while we lose information about the specific mathematical strategy, we gain information about the comprehensive solution process. We argue that the cognitive processes that accumulate during the solving of WPs are reflected in the overall reading pattern. Though we believe that our results support this assumption, we acknowledge that we lose the possibility of explaining in detail what mathematical strategies led to the reading pattern. We believe that the purpose of this paper cannot be to offer a much simpler solution to the very complex process of WP solving that makes other research obsolete. Rather, it offers a comprehensive view that includes both linguistic and mathematical features of WPs and thereby can open a new perspective on the topic, combining theoretical and methodological aspects of both disciplines.

Moreover, the method of eye tracking offers some inherent limitations. Arguably, the situation to solve the WP on a screen and not on paper and to answer orally might have an impact but was considered necessary for the purpose of better eye movement recording. This problem is known from other eye tracking studies in mathematics education (e.g., Andra et al. 2015). We assume that the fact that students were not able to take notes definitely made the tasks more difficult overall. Nevertheless, we assume that the process of WP solving following the Kintsch and Greeno model (1985) is also appropriate without notetaking. Perhaps, a purely mental solution process might reflect the prototype processes described in the model even more adequately. Mobile eye tracking glasses could offer the methodological possibilities to answer this question in the future, by systematically comparing eye movements during WP solving on a computer screen and on paper.

With a pool of 11 items, our results are only applicable to a very small sample of the various forms of WPs. Moreover, our samples do not include non-proficient readers or a group of low-achieving students. To generalize our findings, additional experiments in different contexts are needed.

In our study, we did not control for participants' prior abilities in reading or mathematics. Both play a key role in WP solving (e.g. Boonen et al. 2016; Daroczy et al. 2015; Leiss et al. 2019). In the current study, we were focusing on the relation between task difficulty, performance, and measures of eye movements. In contrast, we did not specify how prior abilities can explain individual differences in WP solving, neither regarding their mathematical abilities nor their reading abilities. Our data shows that the relation between eye movements and task difficulty differs between individuals. Taking into account these facets of prior knowledge in future studies could help explain the cause for these differences. For now, we found that eye movements can be a valuable tool in observing WP solving, and future research could use that tool to further clarify how individual prior knowledge affects the process of WP solving (see also Leiss et al. 2019).

5.8 Conclusion

The aim of this study was to provide a novel approach for the use of eye tracking in mathematics education by using established global measures of eye movements based on a framework of reading. We believe this theoretical approach to connect models of reading and WP solving was supported by promising empirical results,

as the global measures were indeed associated both with task difficulty and characteristics of the reader. This approach emphasizes the importance of processes of reading for mathematical WP solving and could allow future research to tackle new questions through the method of eye tracking.

References

- Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah: Erlbaum.
- Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*(6), 1490–1528.
- Andersson, U. (2007). The contribution of working memory to children's mathematical word problem solving. *Applied Cognitive Psychology*, *21*, 1201–1216.
- Andr , C., Lindstr m, P., Arzarello, F., Holmqvist, K., Robutti, O., & Sabena, C. (2015). Reading mathematics representations: an eye-tracking study. *International Journal of Science and Mathematics Education*, *13*(Suppl. 2), 237–259.
- Blignaut, P. (2009). Fixation identification: the optimum threshold for a dispersion algorithm. *Attention, perception & psychophysics*, *71*(4), 881–895.
- Boonen, A. J. H., 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.
- Boonen, A. J. H., van der Schoot, M., van Wesel, F., de Vries, M. H., & Jolles, J. (2013). What underlies successful word problem solving? A path analysis in sixth grade students. *Contemporary Educational Psychology*, *38*, 271–279.
- Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language*, *86*, 1–19.
- 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.
- Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*(4), 405–438.
- 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.
- Eccles, J., & Wigfield, A. (1995). In the mind of the actor: the structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, *21*(3), 215–225.
- Eccles, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1985). Self-Perceptions, task perceptions, socializing influences, and the decision to enroll in mathematics. In S. F. Chipman, L. R. Brush & D. M. Wilson (Eds.), *Women and mathematics: balancing the equation* (pp. 95–121). London, New York: Psychology Press.
- Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vision Research*, *41*(12), 1561–1574.
- Goldmann, S. R., & Rakestraw, J. A. (2000). Structural aspects of constructing meaning from text. In M. Kamil (Ed.), *Handbook of reading research* (Vol. III, pp. 311–336).
- Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, *84*(1), 76–84.
- 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.
- Holmqvist, K., Nystr m, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: a comprehensive guide to methods and measures*. Oxford: Oxford University Press.
- Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, *43*(4), 358–390.
- Jim nez, L., & Verschaffel, L. (2014). Development of children's solutions on non-standard arithmetic word problem solving. *Revista de Psicodidactica*, *19*(1), 93–123.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354.

- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*(1), 109–129.
- de Koning, B. B., Boonen, A. J. H., & van der Schoot, M. (2017). The consistency effect in word problem solving is effectively reduced through verbal instruction. *Contemporary Educational Psychology*, *49*, 121–129.
- Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., & Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, *10*, 90–115.
- Lee, W.-K., & Wu, C.-J. (2018). Eye movements in integrating geometric text and figure: scanpaths and given-new effects. *International Journal of Science and Mathematics Education*, *16*(4), 699–714.
- Lee, K., Ng, E. L., & Ng, S. F. (2009). The contributions of working memory and executive functioning to problem representation and solution generation in algebraic word problems. *Journal of Educational Psychology*, *101*(2), 373–387.
- Leiss, D., Plath, J., & Schwippert, K. (2019). Language and mathematics—key factors influencing the comprehension process in reality-based tasks. *Mathematical Thinking and Learning*, *21*(2), 131–153.
- Leiss, D., Schukajlow, S., Blum, W., Messner, R., & Pekrun, R. (2010). The role of the situation model in mathematical modelling—task analyses, student competencies, and teacher interventions. *Journal für Mathematik-Didaktik*, *31*(1), 119–141.
- Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, *81*, 521–531.
- Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks: SAGE.
- Matin, E. (1974). Saccadic suppression: A review. *Psychological Bulletin*, *81*, 899–917.
- Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, *9*(4), 329–389.
- Nesher, P., & Teubal, E. (1975). Verbal cues as an interfering factor in verbal problem solving. *Educational Studies in Mathematics*, *6*, 41–51.
- OECD (2006). PISA Released Items—Mathematics. <http://www.oecd.org/pisa/38709418.pdf>
- OECD (2013a). *PISA 2012 assessment and analytical framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing.
- OECD (2013b). PISA 2012 Released Mathematics Items. <https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf>
- OECD (2014). *PISA 2012 Technical Report*. Paris: OECD.
- Pape, S. J. (2004). Middle school childrens' problem-solving behaviour: A cognitive analysis from a reading comprehension perspective. *Journal for Research in Mathematics Education*, *35*(3), 187–219.
- Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European Journal of Cognitive Psychology*, *16*(1/2), 3–26.
- Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *The Quarterly Journal of Experimental Psychology*, *66*(3), 429–452.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.
- Rayner, K., & Liversedge, S. P. (2011). Linguistic and cognitive influences on eye movements during reading. In S. P. Liversedge, I. D. Gilchrist & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 751–766). Oxford: Oxford University Press.
- Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255.
- Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *Psychology of reading* (2nd edn.). New York: Psychology Press.
- Rayner, K., Warren, T., Juhasz, B., & Liversedge, S. P. (2004). The effects of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1290–1301.
- Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157.
- Rellinger, E., Borkowski, J. G., Turner, L. A., & Hale, C. A. (1995). Perceived task difficulty and intelligence: Determinants of strategy use and recall. *Intelligence*, *20*(2), 125–143.
- Reusser, K. (1990). From text to situation to equation: cognitive simulation of understanding and solving mathematical word problems. In H. Mandl, E. De Corte, N. S. Bennett & H. F. Friedrich (Eds.), *Learning and instruction in an international context* (pp. 477–498). New York: Pergamon.

- Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). Die Erfassung des Flow-Erlebens. In J. Stiensmeier-Pelster & F. Rheinberg (Eds.), *Diagnostik von Motivation und Selbstkonzept* (pp. 261–279). Göttingen: Hogrefe.
- Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. H. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York: Academic Press.
- Salvucci, D. D., & Goldberg, J. H. (2000). *Identifying Fixations and Saccades in Eye-Tracking Protocols*. Proceedings of the Eye Tracking Research and Applications Symposium. (pp. 71–78). New York: ACM Press.
- Schindler, M., & Lilienthal, A. J. (2019). Domain-specific interpretation of eye tracking data: towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics*, *101*, 123.
- van der Schoot, M., Bakker Arkema, A. H., Horsley, T. M., & van Lieshout, E. C. D. M. (2009). The consistency effect depends on markedness in less successful but not successful problem solvers: An eye movement study in primary school children. *Contemporary Educational Psychology*, *34*, 58–66.
- SMI. (2011). BeGaze Manual: Version 3.0.
- Verschaffel, L., De Corte, E., & Pauwels, A. (1992). Solving compare problems: An eye movement test of Lewis and Mayer's Consistency Hypothesis. *Journal of Educational Psychology*, *84*(1), 85–94.
- Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Lisse: Swets & Zeitlinger.
- Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole number concepts and operations. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 557–628). Charlotte: Information Age Publishing.
- Vicente, S., Orrantia, J., & Verschaffel, L. (2007). Influence of situational rewording and conceptual rewording on word problem solving. *British Journal of Educational Psychology*, *77*, 829–848.