NONLINEAR DYNAMICS OF READING

AND TEXT COMPREHENSION

Von der Fakultät Bildung

der Leuphana Universität Lüneburg zur Erlangung des Grades

Doktorin der Psychologie

– Dr. phil. –

genehmigte Dissertation von

Monika Tschense

geboren am 14.11.1992 in Wolfen

Eingereicht am: 18.09.2023

Mündliche Verteidigung am: 27.02.2024

Erstbetreuer/ Erstgutachter: Prof. Dr. Sebastian Wallot
Leuphana-Universität Lüneburg

Zweitgutachter: Prof. Dr. Tobias Richter
Julius-Maximilians-Universität Würzburg

Drittgutachter: Prof. Dr. Alexander Freund
Leuphana-Universität Lüneburg

Die einzelnen Beiträge des kumulativen Dissertationsvorhabens sind oder werden ggf. inkl. des Rahmenpapiers wie folgt veröffentlicht:

Tschense, M., & Wallot, S. (2022a). Using measures of reading time regularity (RTR) to quantify eye movement dynamics, and how they are shaped by linguistic information. *Journal of Vision, 22*(6):9, 1–21. https://doi.org/10.1167/jov.22.6.9

Tschense, M., & Wallot, S. (2022b). Modeling items for text comprehension assessment using confirmatory factor analysis. *Frontiers in Psychology*, *13*:966347. https://doi.org/10.3389/fpsyg.2022.966347

Tschense, M., & Wallot, S. (2023). *Using recurrence quantification analysis to measure reading comprehension for long, connected texts* [Manuscript submitted for publication]. Institute for Sustainability Psychology, Leuphana University of Lüneburg.

Veröffentlichungsjahr: 2024

CONTENTS

CHAPTER I: GENERAL RATIONALE

When we are around six years old, we are taught to read. What initially is a slow and effortful endeavor, with practice, turns into a rather fast and automatized process. As adults, we read about 250 words per minute (Brysbaert, 2019). Starting in school, the ability to read quickly gains in relevance for our everyday life. Not only does it contribute to an active participation in society, reading skill is also related to educational attainment, as well as academic and professional achievement (Cox et al., 2003; Peng & Kievit, 2020; Pluck, 2018; Whitten et al., 2019). Yet, recent surveys show that for instance in Germany more than 25% of students in fourth grade demonstrate insufficient reading skills (IGLU 2021; Lorenz et al., 2023), and more than 12% of adults meet criteria of functional illiteracy (LEO 2018; Grötluschen et al., 2020).

But what does *reading* actually entail? Generally speaking, reading could be described as the decoding of written language in order to acquire information. While this might seem quite simple and straight-forward, reading is turns out to be a highly dynamic and complex task. In other words, reading is "an elegantly choreographed dance among a number of visual and mental processes" (Rayner et al., 2016, p. 20). Moreover, the components and operations that make up the reading process as a whole interact in nonlinear ways. Depending on the specific reading situation, effects of interacting factors cannot be expected to be cumulative or proportional, instead, they might amplify or attenuate one another, or even cancel each other out (Wallot & Van Orden, 2011).

This thesis investigates natural text reading, pursuing the overarching aim to establish an objective and easy-to-use measure of text comprehension based on eye movement dynamics. This idea is based on the hypothesis of reading time regularity (RTR; Wallot, 2014; 2016), which postulates an informative link between process measures of reading (e.g., reading times, eye movements), and outcomes of the reading process (e.g., fluency, comprehension). To this end, the degree of regularity in a time series is quantified by means of nonlinear methods that are situated in dynamical systems theory. Thus, another aim of this thesis is to explore the suitability of such methods, here, recurrence quantification analysis (RQA) and sample entropy analysis (SampEn), to capture relevant changes in eye movement behavior during text reading. Furthermore, I emphasize the crucial role of comprehension assessment for reading research.

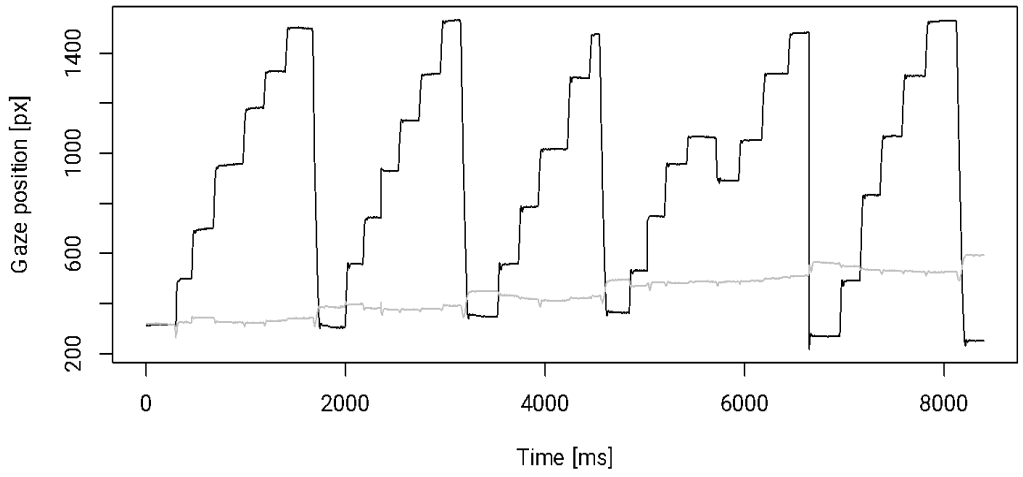
This thesis is structured in four chapters. In the first part of **Chapter I**, the general rationale for the thesis is outlined, and the relevant theoretical and empirical background is summarized. Here, I will briefly describe main components and processes of text reading, and then address two current issues in the area of reading research: the heterogeneity of effects, and the relation between process measures and outcome measures. Subsequently, I will introduce the hypothesis of RTR, present previous studies that underline its advantages, and briefly provide an overview of the two regularity methods employed here. The second part of **Chapter I** features three empirical studies, which systematically investigate the aims of this thesis as outlined above. At first, each study individually will be briefly summarized, and its contribution to the topic of discourse reviewed. Afterwards, the synopsis of all three studies will be discussed considering the limitations of this research, before I conclude with suggestions for future work. **Chapters II**, **III** and **IV** then present the respective research articles.

1 Psychology of Reading: An Overview

Reading a text is a complex task that demands a fine-tuned interplay of motoric, perceptive, and higher-order cognitive processes. Once the reader's attention is directed toward the visual input, their eyes move in a sequence of swift jumps (saccades) alternating with brief pauses (fixations) through the text. Saccades mainly serve the purpose of moving the gaze from one text snippet to the next, whereas fixations allow to process the respective visual input. In addition to saccades and fixations as primary components, gaze behavior during reading is also shaped by regressions to previous parts of the text, and return sweeps from the end of a line to the beginning of the next one (Rayner, 2009; Rayner et al., 2016). When gaze positions are plotted over time, this results in a typical staircase pattern during text reading (see **Figure 1**). Moreover, this illustrates reading as a dynamical task, which requires the rapid, incremental processing of novel information that becomes available with each new fixation.

Figure 1

Staircase pattern of eye movements during text reading



Note. Horizontal (black) and vertical (grey) gaze position are displayed over time. Fixations are intervals of relatively stationary horizontal gaze position; saccades stand out as rapid increase in horizontal gaze position; regressions constitute rapid decrease in horizontal gaze position; in return sweeps the horizontal gaze position drops to the initial level (movement from end of the line to beginning of the line), simultaneously, the vertical gaze position increases (movement from one line to the next).

When and where a reader moves their eyes is largely determined by word recognition (Reichle et al., 2003), which arguably constitutes the most fundamental aspect during reading. The incoming visual information is broken down into features, which are then combined into letters or graphemes, and subsequently assembled into larger orthographic patterns. Finally, the orthographic input can be mapped onto a mental representation of a word in order to access its meaning. These bottom-up processes happen in a parallel and interactive manner, and are highly automatized in typical adult readers. Moreover, they are complemented by top-down information such as word knowledge (i.e., sub-lexical, lexical and semantic properties; Yap & Balota, 2014).

But text reading extends far beyond the recognition of single words, in particular, it requires consecutive words to be parsed into a syntactic structure, and their meaning to be integrated. Usually, these processes are automatic and effortless. But the longer the compositional structures become, the more cognitive effort is required for parsing and interpretation, for instance, in working memory (Lewis et al., 2006). This trade-off is typically investigated within the scope of sentence complexity (e.g., Warren & Gibson, 2002), ambiguity resolution (e.g., MacDonald et al., 1992), or a combination of both (e.g., Kaan & Swaab, 2003; Kim & Christianson, 2013). While researchers still debate the underlying principles of human sentence parsing, there is broad consensus that syntactic and semantic aspects of preceding words usually constrain the integration of upcoming words (Staub, 2014; Staub et al., 2015). Consider examples (1) to (3) below:

(1) Mary likes …
(2) Mary likes her coffee …
(3) Mary likes her coffee with milk and …

Even though the exact continuation is unpredictable in (1), *likes* already demands a direct object that probably has a positive connotation to it; in (2), alternatives are restricted to somehow modify *coffee*, for instance, *black* or *on ice*; the given context in (3) is highly constraining and, given our world knowledge, likely favors the word *sugar*.

The successful parsing and interpretation of a sentence, however, is still not sufficient for text reading. Rather, the overall meaning of a sentence (proposition or information unit) needs to be integrated across sentences in order to attain a representation of the text in memory (Kendeou et al., 2016; Verhoeven & Perfetti, 2008). Due to limited capacities in working memory, not all the information that is encoded in the text can be actively maintained in memory (e.g., Palladino et al., 2001; Radvansky & Copeland, 2001). Thus, abstraction and integration rules apply that transform detailed and local propositions (micro level) into a reduced and more global representation of the text (macro level; Kintsch & van Dijk, 1978; McNamara & Magliano, 2009). At this point, the reader's background knowledge also comes into play, allowing incoming information units to be combined with contents stored in long-term memory. Furthermore, inferences can be drawn that exceed the information explicitly stated in the text, and also knowledge about the text genre, its layout or even the author's motive might be considered by the reader (Kendeou & van den Broek, 2007). The resulting mental model then constitutes a coarse but coherent and situated representation of the text in memory, which reflects basic text comprehension, and serves as a foundation for subsequent tasks (e.g., summarizing, answering questions, discussing, or learning; Kendeou et al., 2016; O'Brien & Cook, 2014).

1.1 Influencing Factors and Heterogeneity of Effects

As outlined above, reading a text is a dynamical and complex task, which consists of multiple interdependent processes that are coordinated across different time scales (e.g., word recognition vs. semantic elaboration) and processing levels (e.g., word vs. sentence vs. text). Therefore, it is not surprising that the majority of studies so far investigated isolated aspects of reading, focusing on a specific manipulation in a controlled set of stimuli. While findings and assumptions generally work well within the scope of the phenomenon of interest, their relevance beyond that is less clear (Rapp & van den Broek, 2005).

An example for this is the transposed letter effect, which describes the relatively intact reading of words despite changes in letter order (e.g., *jugde* vs. *judge*). The effect was demonstrated for several languages, and adopted in models of word recognition (Grainger & Whitney, 2004). However, changes in letter order dramatically disturb reading in Hebrew due to the importance of morphological roots as opposed to the reliance on the orthographic dimension in English (Velan & Frost, 2007). Similarly, effects of lexical features, such as the word frequency effect, are considered to reflect underlying mechanisms of word recognition (Coltheart et al., 2001). The frequency effect refers to facilitated processing of words that often occur in written and spoken language (Brysbaert et al., 2018). Cross-linguistic studies demonstrated this effect for Hebrew and (less strongly) for English, but could not replicate this finding for Serbo-Croatian (Frost & Katz, 1989; Frost et al., 1987). Both examples emphasize the idiosyncratic characteristics of languages on certain aspects of the reading process (Frost, 2012).

In addition to language-driven factors, measures of the reading process are also affected by specific task demands. For instance, Xiong and colleagues (2023) investigated the frequency effect across different reading tasks. Effect size was largest for single-word reading

in a lexical decision task, but reduced by half for sentence reading, and even smaller for word naming. Wallot and colleagues (2013) investigated effects of word frequency and word length during reading of longer texts. They found that lexical variables only explained comparatively little variance in reading times. Moreover, lexical effects decreased systematically with increasing text length. Furthermore, Teng and colleagues (2016) showed that the presence of a frequency effect can entirely depend on the order of reading tasks. While the expected frequency effect was evident when a lexical decision task was performed prior to text reading, the effect vanished when the lexical decision task followed text reading.

Similar results have been found for effects of predictability, demonstrating facilitated word identification given a semantically and syntactically constraining contexts. Typically investigated in sentences, predictability effects have been shown for fixation durations and skipping rates (Abott et al., 2015; Ehrlich & Rayner, 1981), as well as for neural correlates (DeLong et al., 2014; Ito et al., 2016). However, when examining predictability on the scale of entire texts, highly predictable words are rather rare and thus less relevant to overall reading behavior. This was confirmed in a study by Luke and Christianson (2016), in which predictability was investigated in text passages. The authors reported that only 5% of all content words qualified as highly predictable given the prior context.

Varying task demands can also be accompanied by different strategies that are more or less actively employed by the reader. Schotter and colleagues (2014) asked participants to read sentences for overall comprehension or to proofread for spelling errors, and found enhanced frequency effects during proofreading. Interestingly, predictability effects were not modulated by task when the spelling errors resulted in a nonword, but increased for proofreading when spelling errors resulted in unintended, but real words. In a similar design, Andrews and colleagues (2022) demonstrated that predictability metrics interact more

strongly with eye movement measures in proofreading (i.e., longer fixations, less skipping, and more regressions) compared to reading for comprehension. Both studies suggest that readers strategically adapt word processing strategies to accommodate task demands. Furthermore, Andrews and colleagues (2022) found that older readers were more flexible in adjusting their reading strategies to task demands than younger readers.

Even though only a selection of studies was reviewed above, they emphasize that the impact of linguistic properties on process measures of reading significantly depends on idiosyncratic characteristics of languages, specific task demands, reading strategies, and other individual factors. However, such interactions are usually not included in current models of reading, which focus on specific components of the reading process. Consequently, a shift from isolated aspects of reading to shared cognitive operations, allowing for more interactions between text, reader, and situation, might be promising in order to capture the complex dynamics of reading (Rapp & van den Broek, 2005).

1.2 Relation of Reading Process and Reading Outcome

Traditionally, research focused on either the cognitive components and processes involved in reading (e.g., word recognition, sentence parsing, generating inferences) or their aggregation in the sense of a general reading ability or skill (e.g., speed, fluency, comprehension). This separation partly seems to reflect the different perspectives of disciplines (i.e., linguistics and cognitive psychology vs. education and developmental psychology), depicting two sides of the same coin. In the following, I will summarize the body of research that, so far, has investigated the relation between both the reading process and the reading outcome.

As mentioned before, some theories and models of reading emphasize the importance of the word unit, and even consider word processing as a kind of online measure of reading

(Perfetti & Stafura, 2014). Indeed, word decoding skills (i.e., the speed of word identification) and word knowledge (i.e., vocabulary) have been shown to be strong predictors of reading comprehension in children (Carlson et al., 2013) and struggling adult readers (Tighe & Schatschneider, 2016). For proficient adult readers, word knowledge remained a good predictor of comprehension, whereas word decoding only accounted for a small portion of variance (Macaruso & Shankweiler, 2010; Landi, 2010) or did not relate to comprehension at all (Braze et al., 2007). Notably, all studies assessed word decoding, word knowledge, and reading comprehension separately, with unrelated materials. Furthermore, this line of research has primarily focused on effects of individual differences, and is less concerned with establishing an overarching predictive measure for text comprehension.

In contrast to naming or reading speed, eye movements provide a more direct measure of cognitive processing during reading (Rayner & Reingold, 2015). As such, eye movements reflect disruptions in the reading process, manifested in longer fixation durations, shorter saccades, and more regressions. This has been demonstrated for manipulations of linguistic properties of words, sentences, or texts, for instance, for frequency and predictability (Schotter et al., 2014), sentence complexity (Staub, 2010), semantic and structural ambiguity (Sturt, 2007), and passage difficulty (Rayner, Chace, et al., 2006). Further studies have detected group differences in eye movements with respect to overall reading skills, such as children compared to adults (Reichle et al., 2013), younger and older adults (Rayner, Reichle, et al., 2006), or dyslexic versus healthy adult readers (Jones et al., 2007). These results are typically interpreted in terms of a hampered comprehension – even though comprehension was not explicitly assessed in any of these studies.

Only a few studies have so far investigated the relation of process and outcome measures based on the same text stimuli. In a self-paced sentence reading task, Schroeder

(2011) established a positive correlation between comprehension levels and various measures pertaining to reading speed. Southwell and colleagues (2020) observed that enhanced text comprehension was linked to an increased number of (shorter) fixations across data of three experiments. Recently, Mézière and colleagues (2023a; 2023b) corroborated these findings in principle, although the predictive strength of eye movements varied remarkably across different reading tasks and comprehension assessments.

Conversely, a number of prior studies (LeVasseur et al., 2006; 2008; Wallot et al., 2014; 2015) did not succeed in establishing a significant relationship between reading process indicators and reading comprehension, as evidenced in the literature. A critical point that might explain some of the heterogeneous findings above is how reading comprehension was assessed in the respective studies. Since this is the research focus of **Study 2**, I will merely refer to the relevant parts of the thesis here.

2 Reading Time Regularity

Drawing on both, the variability with which linguistic features relate to measures of the reading process, and the research gap between process and outcome of reading, Wallot (2014, 2016) proposed the hypothesis of reading time regularity (RTR). RTR states that the coupling between text (i.e., relevant linguistic information) and reader (i.e., perceptuo-cognitive processes) is reflected in measures of the reading process (e.g., reading times or eye movements). Furthermore, RTR posits that the degree of structure or regularity exhibited in a time series of the reading process can be used to quantify this coupling relation. Hence, efficient and functional coupling should be captured by high degrees of regularity. In turn, variations on either part (e.g., due to manipulations of text features or task demands) should be reflected in changes in regularity. Consequently, RTR can be assumed to be informative about processing difficulty during reading, and thus, related to text comprehension and reading fluency. Tschense and Wallot (2022a) specified the following assumptions underlying RTR:

(A1) Any observable that can be used to measure the reading process (e.g., eye movements) is inherently a random variable.
(A2) When this variable is measured in a reading situation, its values become contingent on relevant properties of the text (e.g., fixations durations become correlated with lexical word properties).
(A3) Because texts are inherently hierarchically ordered sequences (e.g., from words to sentences to text), a random variable that becomes contingent on this sequence will exhibit increased order.
(A4) The coupling between text and reader depends on reading skill and comprehension, thus, efficient coupling implies higher degrees of regularity.

Importantly, RTR infers the coupling strength between text and reader solely based on the degree of regularity present in the process measure. This means that it is not necessary to relate certain text features to changes in reading process measures, but it can be inferred from the relative regularity of the time series. Consequently, RTR renders assumptions about particular effects of linguistic text properties or their interaction with several other factors obsolete. Instead, it can be presumed that a specific reading situation leads to some kind of coupling between the relevant linguistic information and perceptuo-cognitive processes, and this coupling relation can then be quantified. This particularity further qualifies RTR as a potential cross-linguistic measure of the reading process irrespective of idiosyncratic properties of different writing systems.

In the scope of RTR, regularity refers to autocorrelation properties of a time series. Thus, all methods that quantify order in a time series are generally suited to the operationalization of RTR. While there are other methods to be considered (e.g., fractal analysis: Van Orden et al., 2003), this thesis will focus on the application of recurrence quantification analysis (RQA) and sample entropy analysis (SampEn). As the name already implies, RQA calculates the recurrence or repetition of a certain event over time, which represents moments of similar patterns of behavior in a time series. Such instances of recurrent behavior can be visualized and quantified by means of recurrence plots. Different measures can then be derived based on the amount and the clustering of recurrent points within a time series (Marwan et al., 2007; Wallot, 2017). SampEn quantifies the degree of predictability of a time series based on how often patterns of increasing length are repeated within a specified radius in the time series. Time series that are highly deterministic yield a SampEn close to 0; the noisier a time series, the larger is the resulting SampEn (Richman &

Moorman, 2000). Following the assumptions of RTR, closer coupling can thus be inferred from larger recurrence measures, and smaller SampEn.

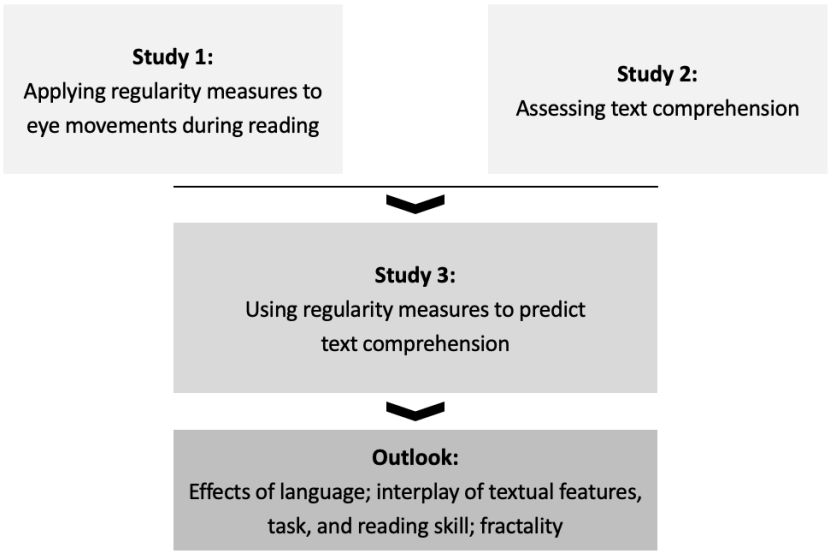
While the concept of RTR has not been explicitly tested so far, some studies that applied measures of regularity to reading tasks have shown promising results. Wijnants and colleagues (2012) compared response times in a naming task of beginner readers with and without dyslexia. Recurrence measures were reduced for dyslexic readers, suggesting less stable reading dynamics. Moreover, recurrence measures correlated positively with reading speed. Furthermore, recurrence measures of self-paced reading times have been shown to be predictive of text comprehension across silent reading and reading aloud, and even more so than reading speed (O'Brien et al., 2014; O'Brien & Wallot., 2016; Wallot et al., 2014). In conclusion, while there is initial evidence for the validity of RTR as an indicator of comprehension and fluency during reading, the assumptions underlying RTR had yet to be systematically tested, and further empirical evidence must be collected.

3 Aims and Contributions

This thesis investigates if the hypothesis of RTR can be utilized to reflect the complex dynamics of text reading. To this end, three empirical studies were conducted, which systematically address this research question (**Figure 2**). **Study 1** examines whether RTR reliably captures the availability of linguistic information, which is a core assumption of the hypothesis. **Study 2** explores the dimensionality of comprehension when assessed after text reading. Finally, **Study 3** evaluates whether RTR can be used as a general means to predict text comprehension, which is the overall motivation for the formulation of RTR. In the following, I will briefly summarize each of these studies and explain their contribution to the overall research question. Afterwards, I will discuss theoretical and practical implications of the empirical findings, and then conclude this first chapter with suggestions for future research. The following **Chapters II**, **III** and **IV** provide the respective research articles.

Figure 2

Schematic Outline (Structure and Aims)



3.1 Study 1: Regularity Measures to Quantify Eye Movement Dynamics (Tschense & Wallot, 2022a)

The first study provided a test for the very basic assumption of RTR, which states that process measures during reading become contingent on relevant linguistic information. In particular, it investigated if eye movement behavior during reading exhibits more regularity than eye movement behavior that is unrelated to reading. To test this, participants' eye movements were recorded during reading (-related) conditions (reading of normal text, shuffled texts, and text grids), and baseline conditions unrelated to reading (looking at fixation cross, blank screen, and randomly distributed circles). The results are in line with assumptions of RTR: Reading and related conditions yielded higher recurrence measures compared to baseline conditions. In a second experiment, participants were uniformly instructed to look at randomly distributed circles, text grids, shuffled texts and normal texts. Additionally, a nonword condition was implemented, to reduce the gap in available linguistic information between text grids and shuffled text. The results suggested that enhanced degrees of linguistic information were reflected in increased recurrence measures.

SampEn also proved to be a sensitive measure of regularity. However, it unexpectedly behaved more similarly to recurrence measures, that is, increased SampEn indicated an enhanced availability of linguistic information. While there were some suggestions that SampEn should not be considered a measure of entropy per se due to its calculation (Porta et al., 2001), an interpretation in terms of RTR seems far-fetched. Nevertheless, it illustrates the necessity of a secondary aim of this study, which was to explore different measures of regularity and their suitability to capture RTR. The same applies to the investigated process measures, since this was the first study investigating eye movements during reading with regularity measures. Here, gaze steps (fluctuations of gaze positions) turned out to be better

suited to capture relevant eye movement dynamics than more aggregated measures such as fixation durations.

3.2 Study 2: Assessment of Text Comprehension (Tschense & Wallot, 2022b)

Reading serves the purpose of extracting information from text, thereby striving for the primary goal of text comprehension. As outlined above, understanding a text is based on its mental representation in memory. Models of discourse representation suggest different processing stages within the generation of such a memory representation (e.g., McNamara & Magliano, 2009; O'Brien & Cook, 2014). One aim of this study was to investigate whether, how, and to what extent three broadly assumed building blocks of mental text representation (i.e., information units at a local level, global level, and inferences) contribute to text comprehension. After reviewing studies that investigated the relation between reading process measures and comprehension, two points stood out. First, there was no gold standard on how to reliably assess text comprehension in terms of the quantity of items or tasks (e.g., summaries, true-/false-statements, open-ended questions). Second, most studies relied on one-shot items without any pre-testing for comprehensibility, difficulty or consistency. Thus, another aim of this study was to investigate how well different items captured comprehension.

To this end, 400 participants read one of three short stories and were then asked to summarize the text, answer open-ended wh-questions, and evaluate yes-/no-statements. The items were constructed to reflect contents on local and global levels of processing, as well as inferred contents. Results of a confirmatory factor analysis yielded evidence that items did not conform to a uni-dimensional concept of comprehension, but rather reflected three distinct,

yet interrelated levels of comprehension. This was true across texts and item types. Regarding the second aim, the initial item pool was already reduced by half based on the judgements of three raters. Data collection began with 16 wh-questions, 60 yes-/no-statements, and 16 main contents per story. After data collection, items with bad psychometric properties were discarded, and further items were excluded based on factor loadings until the models converged. Ultimately, a maximum of ten wh-questions, twelve yes-/no-statements and eight main contents per text were kept.

These results question the common practice of comprehension assessment. On the one hand, it is unclear which specific factors and operational levels of comprehension were assessed in previous studies. Moreover, the use of items that are largely based on experimenters' intuition rather than any theory, pre-testing or post-hoc quality control could falsify the validity of comprehension scores. Both aspects might be crucial to studies that explored the relation between process measures of reading and text comprehension.

3.3 Study 3: Regularity Measures to Predict Text Comprehension (Tschense & Wallot, 2023)

Finally, the third study investigated the assumption of RTR that measures of regularity – reflecting the coupling of text and reader – might be informative about outcome measures of reading. Drawing on findings from the previous two studies, the predictive relation between regularity measures of gaze steps and participants' overall comprehension was explored, using the carefully selected text stimuli and comprehension items from **Study 2**. In order to manipulate coupling strength and, by extension, text comprehension within individuals, participants read the short stories at a comfortable pace (normal reading), as fast as possible

(reading for speed), and as thoroughly as possible (reading for accuracy) while their eye movements were recorded. After each text, participants answered comprehension items.

The results confirmed that recurrence measures of gaze steps are predictive of text comprehension (i.e., comprehension scores normalized by reading speed). This finding was limited to a subset of recurrence measures, which was modulated by the type of comprehension item (i.e., a larger subset for wh-questions, only one measure for yes-/no-statements), and reading condition (i.e., only a main effect of regularity for wh-questions, but an interaction of regularity and reading condition for yes-/no-statements). SampEn did not predict text comprehension. However, the direction of the prediction ran contrary to assumptions of RTR: Higher regularity coincided with less comprehension. A similar prediction effect could be found for the number of fixations, which contradicts the pattern of effects previously demonstrated by Southwell and colleagues (2020).

While in principle the results of this study confirmed the predictive relationship between regularity in eye movement behavior and text comprehension, the negative relation is rather surprising. In a straightforward interpretation of RTR, higher levels of regularity are thought to be associated with better text comprehension due to the need to integrate information units across different scales of text (Kintsch & van Dijk, 1978; Graesser & McNamara, 2011). Hence, further research is needed to better understand the role of regularity as an indicator for reading outcomes. Nevertheless, it was demonstrated that RTR complements more traditional approaches relying on fixation-based measures of the reading process.

4 Discussion

The research presented in the scope of this thesis extends the empirical findings in the context of reading time regularity (RTR). **Study 1** demonstrated that regularity in eye movements is contingent on stimulus-specific information (here, the degree of available linguistic information). The relevance of **Study 2** emerges in light of the heterogeneous body of research concerning the relationship between process measures and outcome measures of reading. As the results suggest, the way in which text comprehension is assessed in a specific study is an important piece of the puzzle. In addition, **Study 3** showed that higher-cognitive processes like text comprehension can be predicted by means of regularity. However, the relation between regularity measures and text comprehension was less straightforward than proposed by RTR.

The complex pattern of results might be partially due to specific task demands imposed by the reading conditions (normal reading vs. reading for speed vs. reading for accuracy), which might not only have modulated text comprehension, but also the strategies employed by the reader (Andrews et al., 2022; Schotter et al., 2014). For instance, Blohm and colleagues (2021) demonstrated that task instructions given to readers influenced their cognitive adjustment before encountering any stimulus. Moreover, the different item types used for comprehension assessment seemed to work differently, which could reflect wh-questions being more complex and difficult to answer. Van Dyke (2021) emphasizes this point from another perspective, noting that all post-hoc assessments require an understanding of the task itself, as well as a comparison of task and text representation. These steps rely on other kinds of executive functions, memory, and even problem-solving skills beyond the reading process itself.

4.1 Limitations and Outlook

Despite the carefully considered experimental designs, there were some limitations to the series of studies included in this thesis. First of all, the measures deployed here originate from complex systems theory. While RQA has been used in numerous cross-disciplinary studies to investigate a wide range of complex phenomena (e.g., http://www.recurrence-plot.tk/bibliography.php), we are only beginning to understand its application to complex behaviors in psychology. Thus, it can only be speculated at this point why recurrence measures worked better for eye movement behavior during reading than SampEn. A possible explanation could be the relation of SampEn to the component of flexibility, that is, the quick and successful adaptation of behavior to changes in a specific situation (Riley & Turvey, 2002; Ward et al., 2018). Following this reasoning, fractal measures could also be of interest in describing such behavioral transitions (Kelty-Stephen & Wallot, 2017).

Furthermore, the results presented within this thesis have to be interpreted in consideration of the specific reading situation that was created due to experimental manipulations. This refers again to the different reading conditions and types of comprehension items, but also to the materials used here. In **Study 1**, only short newspaper articles of 250 words were presented. These texts were not only significantly shorter than the texts used in **Studies 2** and **3**, but newspaper articles also exhibit a different writing style and belong to non-fictional category of texts. In contrast, the short stories that were used in **Studies 2** and **3** had a length of 2,500 words and were fictional texts. Especially the differences in text genre are not trivial, since factors such as vocabulary, familiarity, and relevant previous knowledge are of importance for reading and comprehending a text (e.g., Van Dyke, 2021).

Moreover, measures of RTR emphasize the dynamics of reading. However, the resulting recurrence measures and SampEn were summarized into global scores across the

whole text and reading process. A possibility to tackle this issue would be a more dynamic assessment of comprehension. For instance, comprehension could be assessed after each paragraph or page of text, and then later linked to this specific part. This would also allow for more flexible manipulations of text difficulty.

Recent research has shown that many models of reading do not easily generalize to other languages (Frost, 2012; Frost & Katz, 1989; Velan & Frost 2007). One benefit of RTR is that it is solely based on reading process measures (e.g., response times or eye movements). Thus, in principle, it can be used to assess reading performance independently of the specific linguistic text features or distinct aspects of different writing systems. Following this thought, RTR would be well-suited to assess aspects of reading fluency and text comprehension in cross-linguistic studies.

4.2 Concluding Remarks

Although the theoretical and empirical insights from the presented research contribute in important ways to a better understanding of RTR, it can only be regarded as a first step. Certainly, more systematic research is required to address generalizability across measurements (e.g., reading times vs. gaze positions), variations due to reading tasks (e.g., word-by-word reading vs. natural reading; Teng et al., 2016), reading ability (e.g., healthy vs. dyslexic readers; Wijnants et al., 2012), writing systems (Frost, 2012), as well as processes of flexibility and adaptability of the reader (Kelty-Stephen & Wallot, 2017; Ward et al., 2018).

References

Abbott, M. J., Angele, B., Ahn, Y. D., & Rayner, K. (2015). Skipping syntactically illegal the previews: The role of predictability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1703–1714. http://dx.doi.org/10.1037/xlm0000142

Andrews, S., Veldre, A., Wong, R., Yu, L., & Reichle, E. D. (2022). How do task demands and aging affect lexical prediction during online reading of natural texts? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(3), 407–430. https://doi.org/10.1037/xlm0001200

Blohm, S., Schlesewsky, M., Menninghaus, W., & Scharinger, M. (2021). Text type attribution modulates pre-stimulus alpha power in sentence reading. *Brain and Language, 214*, 104894. https://doi.org/10.1016/j.bandl.2020.104894

Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking up for vocabulary: Reading skill differences in young adults. *Journal of Learning Disabilities, 40*, 226–243. https://doi.org/10.1177/00222194070400030401

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language, 109*, 104047. https://doi.org/10.1016/j.jml.2019.104047

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science, 27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Carlson, E., Jenkins, F., Li, T., & Brownell, M. (2013). The interactions of vocabulary, phonemic awareness, decoding, and reading comprehension. *The Journal of Educational Research, 106*(2), 120–131. https://doi.org/10.1080/00220671.2012.687791

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204–256. https://doi.org/10.1037/0033-295X.108.1.204

Cox, S. R., Friesner, D. L., & Khayum, M. (2003). Do reading skills courses help underprepared readers achieve academic success in college? *Journal of College Reading and Learning, 33*(2), 170-196. https://doi.org/10.1080/10790195.2003.10850147

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass, 8*(12), 631–645. https://doi.org/10.1111/lnc3.12093

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*(5), 263–279. https://doi.org/10.1017/S0140525X11001841

Frost, R., & Katz, L. (1989). Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory & Cognition, 17*(3), 302–310. https://doi.org/10.3758/BF03198468

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13*(1), 104–115. https://doi.org/10.1037/0096-1523.13.1.104

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Grainger, J., & Whitney, C. (2004). Does the huamn mnid raed wrods as a wlohe? *Trends in Cognitive Sciences, 8*(2), 58–59. https://doi.org/10.1016/j.tics.2003.11.006

Grotlüschen, A., Buddeberg, K., Lutz, G., Heilmann, L., & Stammer, C. (2020). Hauptergebnisse und Einordnung zur LEO-Studie 2018 – Leben mit geringer Literalität. In A. Grotlüschen & K. Buddeberg (eds.), *LEO 2018: Leben mit geringer Literalität* (pp. 13–64). wbv. https://dx.doi.org/10.3278/6004740w

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157–171. https://doi.org/10.1016/j.jml.2015.10.007

Jones, M. W., Kelly, M. L., & Corley, M. (2007). Adult dyslexic readers do not demonstrate regularity effects in sentence processing: Evidence from eye-movements. *Reading and Writing, 20*, 933–943. https://doi.org/10.1007/s11145-007-9060-3

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience, 15*(1), 98–110. https://doi.org/10.1162/089892903321107855

Kelty-Stephen, D. G., & Wallot, S. (2017). Multifractality versus (mono-) fractality as evidence of nonlinear interactions across timescales: Disentangling the belief in nonlinearity from the diagnosis of nonlinearity in empirical data. *Ecological Psychology, 29*(4), 259–299. https://doi.org/10.1080/10407413.2017.1368355

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: core components and processes. *Policy Insights from the Behavioral and Brain Sciences, 3*, 62–69. https://doi.org/10.1177/2372732215624707

Kendeou, P., & van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition 35*, 1567–1577. https://doi.org/10.3758/BF03193491

Kim, J. H., & Christianson, K. (2013). Sentence complexity and working memory effects in ambiguity resolution. *Journal of Psycholinguistic Research, 42*, 393–411. https://doi.org/10.1007/s10936-012-9224-4

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Landi, N. (2010). An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Reading and Writing, 23*, 701–717. https://doi.org/10.1007/s11145-009-9180-z

LeVasseur, V. M., Macaruso, P., Palumbo, L. C., & Shankweiler, D. (2006). Syntactically cued text facilitates oral reading fluency in developing readers. *Applied Psycholinguistics, 27*(3), 423–445. https://doi.org/10.1017/S0142716406060346

LeVasseur, V. M., Macaruso, P., & Shankweiler, D. (2008). Promotion gains in reading fluency: A comparison of three approaches. Reading and Writing, 21, 205–230. https://doi.org/10.1007/s11145-007-9070-1

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*(10), 447–454. https://doi.org/10.1016/j.tics.2006.08.007

Lorenz, R., McElvany, N., Schilcher, A., & Ludewig, U. (2023). Lesekompetenz von Viertklässlerinnen und Viertklässlern im internationalen Vergleich: Testkonzeption und Ergebnisse von IGLU 2021. In N. McElvany, R. Lorenz, A. Frey, F. Goldhammer, A. Schilcher & T. C. Stube (eds), *Lesekompetenz von Grundschulkindern im internationalen Vergleich und im Trend über 20 Jahre* (pp. 53–87). Waxmann.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology, 24*(1), 56–98. https://doi.org/10.1016/0010-0285(92)90003-K

Macaruso, P., & Shankweiler, D. (2010). Expanding the simple view of reading in accounting for reading skills in community college students. *Reading Psychology, 31*(5), 454–471. https://doi.org/10.1080/02702710903241363

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports, 438*, 237–329. https://doi.org/10.1016/j.physrep.2006.11.001

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation, 51*, 297–384. https://doi.org/10.1016/S0079-7421(09)51009-2

Mézière, D. C., Yu, L., McArthur, G., Reichle, E. D., & von der Malsburg, T. (2023a). Scanpath regularity as an index of reading comprehension. *Scientific Studies of Reading*, 1–22. https://doi.org/10.1080/10888438.2023.2232063

Mézière, D. C., Yu, L., Reichle, E. D., Von Der Malsburg, T., & McArthur, G. (2023b). Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly, 58*, 425–449. https://doi.org/10.1002/rrq.498

O'Brien, E., & Cook, A. E. (2014). Models of discourse comprehension. In A. Pollatsek & R. Treiman (eds.), *The Oxford Handbook of Reading* (online ed., pp. 217–231). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324576.013.16

O'Brien, B. A., & Wallot, S. (2016). Silent reading fluency and comprehension in bilingual children. *Frontiers in Psychology, 7*, 1265. https://doi.org/10.3389/fpsyg.2016.01265

O'Brien, B. A., Wallot, S., Haussmann, A., & Kloos, H. (2014). Using complexity metrics to assess silent reading fluency: A cross-sectional study comparing oral and silent reading. *Scientific Studies of Reading, 18*(4), 235–254. https://doi.org/10.1080/10888438.2013.862248

Palladino, P., Cornoldi, C., De Beni, R., Pazzaglia, F. (2001). Working memory and updating processes in reading comprehension. *Memory & Cognition 29*, 344–354. https://doi.org/10.3758/BF03194929

Peng, P. and Kievit, R.A. (2020), The Development of Academic Achievement and Cognitive Abilities: A Bidirectional Perspective. *Child Development Perspectives, 14*(1), 15–20. https://doi.org/10.1111/cdep.12352

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading, 18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Pluck, G. (2018). Lexical reading ability predicts academic achievement at university level. *Cognition, Brain, Behavior: An Interdisciplinary Journal, 22*(3), 175–196. https://doi.org/10.24193/cbb.2018.22.12

Radvansky, G.A., Copeland, D.E. (2001). Working memory and situation model updating. *Memory & Cognition 29*, 1073–1080. https://doi.org/10.3758/BF03206375

Rapp, D. N., & Van Den Broek, P. (2005). Dynamic text comprehension: An integrative view of reading. *Current Directions in Psychological Science, 14*(5), 276–279. https://doi.org/10.1111/j.0963-7214.2005.00380.x

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

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. https://doi.org/10.1207/s1532799xssr1003_3

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging, 21*(3), 448–465. https://doi.org/10.1037/0882-7974.21.3.448

Rayner, K., & Reingold, E. M. (2015). Evidence for direct cognitive control of fixation durations during reading. *Current Opinion in Behavioral Sciences, 1*, 107–112. https://doi.org/10.1016/j.cobeha.2014.10.008

Rayner, K., Schotter, E. R., Masson, M. E., Potter, M. C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest, 17*(1), 4–34. https://doi.org/10.1177/1529100615623267

Reichle, E. D., Liversedge, S. P., Drieghe, D., Blythe, H. I., Joseph, H. S., White, S. J., & Rayner, K. (2013). Using EZ Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review, 33*(2), 110–149. https://doi.org/10.1016/j.dr.2013.03.001

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445–476. https://doi.org/10.1017/S0140525X03000104

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology – Heart and Circulatory Physiology, 278*(6), H2039–H2049. https://doi.org/10.1152/ajpheart.2000.278.6.H2039

Riley, M. A., & Turvey, M. T. (2002). Variability and determinism in motor behavior. *Journal of Motor Behavior, 34*(2), 99–125. https://doi.org/10.1080/00222890209601934

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition, 131*(1), 1–27. https://doi.org/10.1016/j.cognition.2013.11.018

Schroeder, S. (2011). What readers have and do: Effects of students' verbal ability and reading time components on comprehension with and without text availability. *Journal of Educational Psychology, 103*(4), 877–896. https://doi.org/10.1037/a0023731

Southwell, R., Gregg, J., Bixler, R., & D'Mello, S. K. (2020). What eye movements reveal about later comprehension of long connected texts. *Cognitive Science, 44*(10), e12905. https://doi.org/10.1111/cogs.12905

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. Cognition, 116(1), 71–86. https://doi.org/10.1016/j.cognition.2010.04.002

Staub, A. (2014). Reading sentences: Syntactic parsing and semantic interpretation. In A. Pollatsek & R. Treiman (eds.), *The Oxford Handbook of Reading* (online ed., pp. 202–216). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324576.013.15

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition, 105*(2), 477–488. https://doi.org/10.1016/j.cognition.2006.10.009

Teng, D. W., Wallot, S., & Kelty-Stephen, D. G. (2016). Single-word recognition need not depend on single-word features: Narrative coherence counteracts effects of single-word features

that lexical decision emphasizes. *Journal of Psycholinguistic Research, 45*(6), 1451–1472. https://doi.org/10.1007/s10936-016-9416-4

Tighe, E. L., & Schatschneider, C. (2016). Examining the relationships of component reading skills to reading comprehension in struggling adult readers: A meta-analysis. *Journal of Learning Disabilities, 49*(4), 395–409. https://doi.org/10.1177/0022219414555415

Tschense, M., & Wallot, S. (2022). Using measures of reading time regularity (RTR) to quantify eye movement dynamics, and how they are shaped by linguistic information. *Journal of Vision, 22*(6):9, 1–21. https://doi.org/10.1167/jov.22.6.9

Tschense, M., & Wallot, S. (2022). Modeling items for text comprehension assessment using confirmatory factor analysis. *Frontiers in Psychology, 13*:966347. https://doi.org/10.3389/fpsyg.2022.966347

Tschense, M., & Wallot, S. (2023). *Using recurrence quantification analysis to measure reading comprehension for long, connected texts* [Manuscript submitted for publication]. Institute for Sustainability Psychology, Leuphana University of Lüneburg.

Van Dyke, J. A. (2021). Introduction to the special issue: Mechanisms of variation in reading comprehension: Processes and products. *Scientific Studies of Reading, 25*(2), 93–103. https://doi.org/10.1080/10888438.2021.1873347

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132*(3), 331–350. https://doi.org/10.1037/0096-3445.132.3.331

Velan, H., Frost, R. (2007) Cambridge University versus Hebrew University: The impact of letter transposition on reading English and Hebrew. *Psychonomic Bulletin & Review, 14*(5), 913–918. https://doi.org/10.3758/BF03194121

Verhoeven, L., & Perfetti, C. (2008). Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology, 22*(3), 293–301. https://doi.org/10.1002/acp.1417

Wallot, S. (2014). From “cracking the orthographic code” to “playing with language”: Toward a usage-based foundation of the reading process. *Frontiers in Psychology, 5*, 891. https://doi.org/10.3389/fpsyg.2014.00891

Wallot, S. (2016). Understanding reading as a form of language-use: A language game hypothesis. *New Ideas in Psychology, 42*, 21–28. https://doi.org/10.1016/j.newideapsych.2015.07.006

Wallot, S. (2017). Recurrence quantification analysis of processes and products of discourse: A tutorial in R. *Discourse Processes, 54*(5–6), 382–405. https://doi.org/10.1080/0163853X.2017.1297921

Wallot, S., Hollis, G., & van Rooij, M. (2013). Connected text reading and differences in text reading fluency in adult readers. *PLoS ONE, 8*(8), e71914. https://doi.org/10.1371/journal.pone.0071914

Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, H., & Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1745–1765. https://doi.org/10.1037/xlm0000030

Wallot, S., & Van Orden, G. (2011). Toward a lifespan metric of reading fluency. *International Journal of Bifurcation and Chaos, 21*(4), 1173–1192. https://doi.org/10.1142/S0218127411028982

Ward, P., Gore, J., Hutton, R., Conway, G. E., & Hoffman, R. R. (2018). Adaptive skill as the conditio sine qua non of expertise. *Journal of Applied Research in Memory and Cognition, 7*(1), 35–50. https://doi.org/10.1016/j.jarmac.2018.01.009

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition, 85*(1), 79–112. https://doi.org/10.1016/S0010-0277(02)00087-2

Whitten, C., Lobby, S., & Sullivan, S. L. (2019). The impact of pleasure reading on academic success. *Journal of Multidisciplinary Graduate Research, 2*(1), 48–64.

Wijnants, M.L., Hasselman, F., Cox, R.F.A., Bosman, A.M.T, & Van Orden, G. (2012) An interaction-dominant perspective on reading fluency and dyslexia. *Annals of Dyslexia, 62*, 100–119. https://doi.org/10.1007/s11881-012-0067-3

Xiong, J., Yu, L., Veldre, A., Reichle, E. D., & Andrews, S. (2023). A multitask comparison of word- and character-frequency effects in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(5), 793–811. https://doi.org/10.1037/xlm0001192

Yap, M. J., & Balota, D. A. (2014). Visual word recognition. In A. Pollatsek & R. Treiman (eds.), *The Oxford Handbook of Reading* (online ed., pp. 26–43). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324576.013.4

CHAPTER II: REGULARITY MEASURES TO QUANTIFY EYE MOVEMENT DYNAMICS

Tschense, M., & Wallot, S. (2022). Using measures of reading time regularity (RTR) to quantify eye movement dynamics, and how they are shaped by linguistic information. *Journal of Vision*, *22*(6):9, 1–21. https://doi.org/10.1167/jov.22.6.9

Authorship status:	First author
Publication status:	Published on May 25, 2022
Scientific journal:	Journal of Vision

The article is presented in its published version.

Journal of Vision (2022) 22(6):9, 1–21

Using measures of reading time regularity (RTR) to quantify eye movement dynamics, and how they are shaped by linguistic information

Monika Tschense

Institute of Psychology, Leuphana University of Lüneburg, Lüneburg, Germany
Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

Sebastian Wallot

Institute of Psychology, Leuphana University of Lüneburg, Lüneburg, Germany
Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

In this article, we present the concept of reading time regularity (RTR) as a measure to capture reading process dynamics. The first study is concerned with examining one of the assumptions of RTR, namely, that process measures of reading, such as eye movement fluctuations and fixation durations, exhibit higher regularity when contingent on sequentially structured information, such as texts. To test this, eye movements of 26 German native speakers were recorded during reading-unrelated and reading-related tasks. To analyze the data, we used recurrence quantification analysis and sample entropy analysis to quantify the degree of temporal structure in time series of gaze steps and fixation durations. The results showed that eye movements become more regular in reading compared to nonreading conditions. These effects were most prominent when calculated on the basis of gaze step data. In a second study, eye movements of 27 native speakers of German were recorded for five conditions with increasing linguistic information. The results replicate the findings of the first study, verifying that these effects are not due to mere differences in task instructions between conditions. Implications for the concept of RTR and for future studies using these metrics in reading research are discussed.

Introduction

What guides the reading process? Reading is a complex cognitive process bringing together perceptual-motoric skills, executive functions, memory capacities, and language knowledge (e.g., Rayner & Reichle, 2010). A general assumption all theories and models of reading share is that the reading process is driven by linguistic features of written language, at least to some extent. This is particularly evident for the front-end processes of reading, such as visual word recognition, where lexical features (e.g., word length, word frequency, semantic properties) substantially impact word reading times (Grainger & Jacobs, 1996; Ziegler et al., 2000). Consequently, it is implemented in more encompassing models of eye movements during reading in which lexical features govern fixation durations and saccadic programming (Engbert et al., 2005; Reichle et al., 2009). Moreover, higher-level theories of reading and models of discourse comprehension assume that linguistic features of a text, such as propositional density, situation model dimensions, and syntactic complexity, drive reading times for connected text (Graesser et al., 2004; Kintsch & Keenan, 1973; Zwaan et al., 1995). This is further supported by studies showing that mind-wandering during reading leads to a detachment of eye movement

https://doi.org/10.1167/jov.22.6.9 Received August 6, 2021; published May 25, 2022 ISSN 1534-7362 Copyright 2022 The Authors

This work is licensed under a Creative Commons Attribution 4.0 International License.

measures from linguistic text features (Faber et al., 2020; Schad et al., 2012). Hence, a basic presumption might be that there is indeed a systematic relationship between linguistic text features and the reading process. Following this line of thought, linguistic features should account for (a large fraction of) the variance of observables of the reading process (e.g., word frequency should unequivocally predict sentence reading times).

However, the coupling between reader performance and linguistic text characteristics strongly varies between individuals (Rayner et al., 2006; Traxler et al., 2012), tasks (Teng et al., 2016; Wallot et al., 2013), and languages (Frost, 2012; Holden & Van Orden, 2002). For example, the effect sizes of word frequency and word length differ substantially between reading tasks presenting isolated words or sentences as compared to reading longer, connected texts. Wallot and colleagues (2014) report smaller effect sizes for connected texts compared to reading tasks that emphasize shorter language segments. Besides, there is evidence that effects of lexical features decrease systematically for reading of connected text (Wallot et al., 2013). Furthermore, such effects can even depend entirely on the order in which reading tasks are performed. As shown by Teng and colleagues (2016), word frequency effects for a lexical decision task disappeared when participants had performed a connected text reading task beforehand, while the frequency effect stayed completely intact when the lexical decision task was performed first.

This variability of results regarding the relationship between text features and measures of the reading process is evident not only across tasks but also across languages (Frost, 2012). So showed Holden and Van Orden (2002) that the strength of the word frequency effect varies rather strongly for different languages. Similarly, reading in many languages has been shown to be quite robust regarding changes in letter order, which has been subsequently described as a core property of reading at the neurophysiological level (Whitney & Cornelissen, 2005). Yet, research shows that changes in letter order pose a great challenge for readers of Hebrew (Velan & Frost, 2007). Taken together, it is clear that text features play an important role in controlling the reading process, but the way they do so is not easy to generalize across reading situations, languages, and readers. This also makes it difficult to build a general theory of the reading process based on text features as its driving factors.

Reading time regularity

We thus introduce the concept of reading time regularity (RTR) as a general means to assess the influence of (linguistic) information on perceptual-cognitive processes during reading (Wallot, 2014, 2016). From the perspective of RTR, a process that has a high degree of regularity is a process that evolves comparatively stable over time. Such a process is not subject to larger perturbations or dampens them out quickly. Perturbations of the reading process usually result in conjunction with problems of concentration (e.g., mind-wandering: Faber et al., 2020; Schad et al., 2012), comprehension and text difficulty (Rayner et al., 2006), reading skill (Reichle et al., 2013), or surprise or failure of prediction (Booth et al., 2018). This means that a reader does not efficiently continue to read but has to integrate information differently, search for information, or change the situation model (McNerney et al., 2011). Such changes are usually evident in the reading time course as reflected in long reading times, increased variability of reading times, or specific eye movements, such as regressions.

If a reader is skilled, he or she will be able to solve such conflicts quickly and restore comprehension, so that misunderstandings do not increase the probability for additional comprehension problems later in the text. Both the quick resolution of such conflicts, as well as the reduced probability of encountering such conflicts, will reduce the variability of reading process measures, such as word reading times or eye movement fluctuation, and hence increase the temporal structure, the regularity of the process measure in question. Accordingly, regularity can be seen as a marker of skilled and efficient reading.

Or course, the basic input for what is efficient reading or reading problems is the linguistic information present in a text, which can span the whole range of sublexical, lexical, semantic, syntactic, and discourse-level features. As we have laid out above, the problem is that the effects of each of these features is highly variable across task, person, and language when trying to relate specific text features to changes in reading process measures, but observables.

Here, we propose that RTR might offer a solution to the problem of the variability with which linguistic features relate to measures of the reading process. As explained above, a reading process of high regularity captures efficient and skilled reading, and accordingly good or at least sufficient comprehension. In order to draw this conclusion, however, we do not need to relate specific text features to changes in reading process measures, but we can simply make such an inference based on the relative degree of regularity.

This also means that we do not need to make particular assumptions about the effect of particular text features in question, or how several of such features might interact to bring about a particular effect, or why such an effect seems to be strong under some reading conditions but weak under others. We can assume that there is a coupling between the relevant linguistic information in a particular instance of reading and the cognitive-perceptual processes involved in reading, and if that coupling is efficient

and functional, this will be marked by a high degree of regularity.

Our proposal rests on the following assumptions:

A1. Any observable that can be used to measure the reading process (e.g., eye movements) is inherently a random variable of sorts.
A2. When this variable is measured in a reading situation, its values become contingent on some properties of the text that are relevant for the reader (e.g., fixations durations become correlated with lexical word properties).
A3. Because texts are inherently hierarchically ordered sequences (e.g., from topic to syntax/word order to lexical—and sublexical—properties), a random variable that becomes contingent on this sequence will exhibit increased order.
A4. Because ability of the reader to couple with a text depends on reading skill and comprehension, efficient coupling implies higher degrees of regularity.

Assumptions A2 and A4 are to some degree restatements of the general assumption shared by all models of reading, namely, that linguistic features co-control the reading process. Importantly, however, in the logic of RTR, linguistic text features are not necessary to describe the coupling between reader and text, but it can be inferred from the degree of regularity of a reading process measure alone.

Statistically, RTR captures the regularity, that is, autocorrelation properties, of process measures. Hence, the degree of RTR of a reading process measure can in principle be calculated by any statistic that captures order of a sequence or time series, such as recurrence quantification analysis (Zbilut & Webber, 1992), or sample entropy analysis (Richman & Moorman, 2000). The fact that RTR is solely based on the values of an observable of the reading process, specifically on their sequential properties, but not particularly on text features, can address the challenges outlined above. This is what distinguishes RTR from other attempts to define cognitive coupling (e.g., Mills et al., 2017). Before summarizing some potential applications of RTR in reading research, we provide a brief description of the regularity measures employed in this study. Further information about the parameter estimation for these measures is provided in the Method section.

Measures of regularity

Recurrence quantification analysis: Recurrence quantification analysis (RQA) can be used to quantify various dynamic properties of a time series related to the degree structure of its temporal evolution. Effectively, the RQA measures we employ here capture different kinds of autocorrelation in a time series. They capture different aspects of clustering of data points over time, which is how, i.e., individual data points forming larger patterns within a time series. This can be visualized by means of recurrence plots (RPs) based on which several complexity measures can be derived quantifying the density of recurrence points and their line structures (Zbilut & Webber, 1992). Several RQA measures can be extracted from an RP, but we will focus on the most common measures—recurrence rate (RR), determinism rate (DET), average diagonal line length (ADL), and maximum diagonal line length (MDL): The RR refers to the density of recurrence points, providing information about the repetitiveness of individual values or coordinates within a time series. The less stochastic and the more deterministic a process is, the more recurrent points occur in connected trajectories as opposed to single recurrence points. How many recurrent points occur in diagonal lines as opposed to individual repetitions is indicated by DET. The line length can also be extracted, either as ADL or as MDL. While these measures can distinguish different dynamics properties in certain systems (Marwan et al., 2007), in data with a strong stochastic component, such as eye movement fluctuations, they are often highly correlated. Accordingly, we aim to investigate whether all or just some of them make good indicators of regularity.

RQA has been applied to a variety of research areas, but it has also been used to analyze reading times from dyslexics and nonimpaired controls during a naming task (Wijnants et al., 2012), as well as text reading times of children and adults (Wallot et al., 2014). These studies report lower RQA measures for dyslexic reading compared to controls and that higher RQA measures correlated positively with reading speed and comprehension, probably reflecting a more skilled and efficient processing of text. In line with these results, higher values of RR, DET, ADL, and MLD indicate higher regularity according to RTR.

Sample entropy analysis: Sample entropy analysis (SampEn) quantifies the degree of predictability of a time series (Richman & Moorman, 2000). It takes into account the number of matching sequences identified within a tolerance band defined by a radius r, excluding self-matches. Specifically, SampEn is the average probability that a sequence with length of $m + 1$ data points finds a matching sequence within r, given that a match for m data points has already been found. Highly periodic, deterministic time series are easily predictable (i.e., if sequences of m points repeat, then sequences of $m + 1$ points are also likely to repeat), yielding a $SampEn = 0$. In contrast, a time series that is very noisy yields a $SampEn > 0$.

While sample entropy has been increasingly employed in sport science and motor control research, it has not yet been used to investigate reading data. As a measure of entropy, higher values of SampEn might

indicate lower regularity in terms of RTR. However, because RTR is not about entropy per se but about how well patterns of different length are contained within each other, SampEn might behave more like an entropy rate measure (Porta et al., 2001). That is a measure of complexity, and as such, SampEn might in fact be higher during reading compared to baseline conditions with fewer external information to be processed.

Potential applications of reading time regularity in reading research

Insofar as some of the measures described above turn out to be a valid metrics for capturing functional coupling of linguistic information and perceptual-cognitive processing, RTR has potential applications for reading research. First of all, RTR might make a suitable measure of reading fluency. While reading fluency is conceived as relatively effortless reading with at least average to good comprehension (O'Brien et al., 2014), it is often operationalized as overall reading speed or reading time components. Here, level of speed is used as a stand-in measure for the reading process, because of the positive correlation between skill and reading speed (Fuchs et al., 2001). However, reading speed during text reading is not always substantially related to comprehension, calling this relationship into question (LeVasseur et al., 2006, 2008; Wallot et al., 2014). Instead of using speed as a key characteristic of the reading process, it can equally be seen as an outcome of reading ability and hence reading fluency instead of being a process per se. So far, this circularity issue constitutes an experimental confound in the presumed positive relationship of reading speed and comprehension, which is difficult to avoid empirically. Moreover, the relationship between reading speed and comprehension is complex: While an increase in reading speed can lead to a decrease in comprehension in a trade-off relationship, it can also lead to increases in comprehension. But speed is also thought to correlate positively with comprehension as a general aspect of reading ability.

Therefore, adding the concept of RTR into an operational definition of reading fluency might be able to resolve this conceptual problem: When RTR is used as a measure for reading process fluency in the sense of an effortful, functional execution of the reading process, speed can be solely treated as an outcome variable—and measures of reading time regularity have shown a predictive link to reading speed and comprehension, as well as capture their trade-off relation very well (Wallot et al., 2014).

Since the calculation of RTR does not depend on specific linguistic text features, it can, in principle, be used as a cross-linguistic measure for the prediction of reading comprehension, irrespective of the particular properties of different writing systems and their consequences for reading.

Prior work using measures of regularity of the reading process has shown that the degree of regularity in reading time data is positively correlated to reading comprehension. Notably, RTR properties reliably predicted text comprehension better than reading speed (O'Brien et al., 2014; O'Brien & Wallot, 2016; Wallot et al., 2014), and preliminary results from an eye tracking study corroborated the power of RTR measures in predicting text comprehension using eye movements over and above standard eye movement features related to comprehension, such as fixation durations, number of fixations, and percentage of regressive eye movements (Wallot et al., 2015).

However, these results were obtained before the formulation of RTR and formed the basis for this concept. No prospective tests of this hypothesis have been performed, and, crucially, none of the assumptions (A1–A4) outlined above have been prospectively tested. Hence, the goal of the current article is to test and investigate the foundational measurement assumptions of RTR, particularly A2 and A3, regarding the basic effect of (linguistic) information on process measures—time series of eye movement records—on measures that capture the regularity of such time series. We will return to the discussion of applications of RTR in reading research at the end of the discussion section.

Experiment 1

In order to test one of the basic assumptions of RTR, namely that the presence of external (linguistic) information leads to an increase in process regularity, we constructed an eye movements experiment. We included six conditions: Three contained little to no visual information, two contained information associated with reading, and one condition contained proper text. Figure 1 illustrates the conditions. Participants' eye movements were subjected to RQA, FA, and SampEn in order to quantify the degree of regularity of eye movements in each of these conditions.

Hypotheses

Drawing on the concept of RTR, it is hypothesized that the presence of external linguistic information (see Figures 1d–f) leads to increases in regularity compared to control conditions that do not contain such information (see Figures 1a–c). This is expected because the coupling between cognitive processing and the sequential structure of that information leads eye movement dynamics to become more regular. This hypothesis is tested using gaze step (Stephen & Mirman,

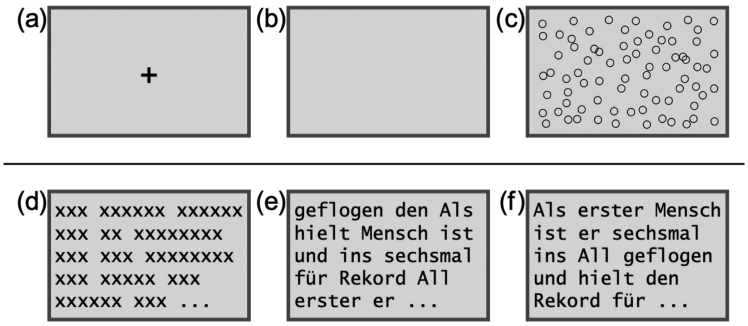


Figure 1. Schematic examples for the experimental conditions. The top panel illustrates the baseline conditions with (a) fixation cross, (b) blank screen, and (c) circles. The bottom panel shows the reading conditions with (d) text grid, (e) shuffled text, and (f) normal text.

2010). The gaze step is the spatial distance between two position measures of the raw eye movement record (see Method section for details on calculating gaze step). This is done because some of the baseline conditions, particularly the fixation cross and blank screen conditions, exhibit drift-like behavior and are not well parsable into fixation durations (Yarbus, 1967/2013) since fixations are largely absent in the respective time series.

In an exploratory part, we will evaluate to what extent the reading conditions (see Figures 1d–f) can be distinguished from one another by means of the described regularity measures. Since normal text provides the maximal degree of linguistic information possible during reading, we predict the text condition to lead to increased regularity in eye movement dynamics compared to text grids and shuffled texts. However, it is currently unknown which of the regularity measures described above capture these differences best—or at all. Analyses will be based on both gaze step data as well as time series of fixation durations extracted for the three reading conditions. A more general aim of this study is also to test several regularity indicators (recurrence and entropy measures) that might be principally suitable for the operationalization of RTR with regard to their validity and sensitivity to distinguish between conditions exhibiting differences regarding their degree of external (linguistic) information.

Method

Participants

Twenty-six native speakers of German with normal or corrected-to-normal vision participated in the study and received a compensation of 15€. One participant terminated the experiment before completion and was therefore discarded from any analysis. Due to technical problems during calibration procedure and data recording, two other participants had to be excluded. Furthermore, data of a fourth participant was excluded due to excessive artifacts and blinks. Thus, the final sample consisted of 22 participants (15 female) with a mean age of 27.63 years ($SD = 9.59$). See Appendix A for further information about the participants. Prior to the experiment, written informed consent was obtained from all participants. The study was approved by the Ethics Council of the Max Planck Society and followed the ethical principles of the Declaration of Helsinki.

Stimuli

The experiment was composed of six conditions, including three conditions unrelated to reading, another two conditions reflecting certain aspects of reading, and one condition consisting of normal text reading (see Figure 1). For the reading-unrelated conditions (baseline conditions), participants were shown (a) a static fixation cross in the middle of the screen, (b) a blank screen, or (c) a screen filled with circles. For the circle condition, 500 circles with black outline at a size of 10 px were randomly distributed on the screen, and a total of seven circle patterns were created.

The other three conditions (reading conditions) consisted of (d) text grids, (e) shuffled texts, or (f) actual newspaper texts. Reading conditions were based on articles from the German daily newspaper *Die Welt* published in January 2018. Chosen articles consisted of 150 to 200 words and did not concern highly divisive topics. For seven newspaper sections (Economics, Feuilleton, Finances, Politics, Science, Society, Sports), two articles each were selected and randomly assigned to one of two lists. Some key descriptive text characteristics are summarized in Table 1. See Appendix B for all characteristics collected.

For conditions (d) and (e), all special characters within a text were removed and all content-dependent or infrequent abbreviations were fully spelled out.

List	Section	Words	Sentences	Words per sentence	Syllables per word	Type frequency		Annotated type frequency	
						M	*SD*	*M*	*SD*
List A	Economics	180	10	18.00	2.08	4.18	1.54	4.06	1.54
	Feuilleton	195	10	19.50	1.92	3.98	1.96	3.87	1.96
	Finances	194	11	17.64	2.30	3.79	1.92	3.66	1.92
	Politics	177	11	16.09	2.29	4.03	1.93	3.81	1.93
	Science	157	9	17.44	2.18	3.80	1.85	3.72	1.85
	Society	201	15	13.40	1.98	4.27	1.57	4.11	1.57
	Sports	189	12	15.75	2.06	3.93	2.02	3.74	2.02
	Overall	184.71	11.14	16.83	2.12	4.00	0.17	3.85	0.17
List B	Economics	197	14	14.07	2.25	3.98	1.94	3.91	1.94
	Feuilleton	162	10	16.20	2.28	3.87	1.92	3.75	1.92
	Finances	197	12	16.42	2.21	3.68	2.06	3.52	2.06
	Politics	187	9	20.78	2.10	4.08	1.74	3.94	1.74
	Science	179	10	17.90	2.04	4.01	1.91	3.82	1.91
	Society	189	14	13.50	2.06	4.09	1.79	3.90	1.79
	Sports	158	7	22.57	2.23	3.88	1.97	3.76	1.97
	Overall	181.29	10.86	17.35	2.17	3.94	0.14	3.80	0.14

Table 1. Text characteristics of the selected newspaper articles for List A and List B. *Notes*: Number of syllables, type frequency, and annotated type frequency were obtained from dlexDB (Heister et al., 2011). Given frequency values are logarithmized.

Subsequently, every letter got replaced by "x", resulting in grid-like structures for condition (d). While text grids reveal certain surface characteristics (e.g., word length), they prohibit any semantic access. For condition (e), a random sequence of words was generated by shuffling the text of the newspaper articles. Thus, a coherent, in-depth processing beyond the individual word semantics was not possible.

Procedure

The study was carried out in a soundproof both with dimmed light. Participants were seated 70 cm in front of an LCD monitor (size: 24 in., refresh rate: 144 Hz, resolution: 1920 × 1080 px). Their head was supported by a head and chin rest to obtain high tracking accuracy. An EyeLink 1000 (SR-Research, Ottawa, Ontario, Canada) was used for monocular data recording of the left eye at a sampling rate of 1000 Hz. Visual stimuli were presented in white on a black background. Fixation cross was presented with 1° visual angle, circle diameter was 0.3° visual angle, and letter width was 0.5° visual angle.

The experiment was conducted in one session that took approximately 90 minutes, depending on participants' individual reading speed. Halfway through the experiment, participants were allowed to take a short break. At the beginning of each half of the experiment, a 12-point calibration with random sequence was performed, followed by a validation of the measured points. A questionnaire succeeded the experiment to gather demographic information.

Participants were randomly assigned to one of two stimulus lists that differed in terms of newspaper articles: Either they were shown Set A, consisting of seven newspaper articles as coherent texts, and or Set B, including the other seven newspaper articles as shuffled texts and text grids, or vice versa. However, texts were selected so that each set contained one article from each of the seven sections of the newspaper (see stimuli above). Participants were presented with seven trials per condition, resulting in a total of 42 trials per participant. The sequence of trials was fully randomized for each participant.

While participants were asked to fixate the fixation cross for (a), they were allowed to look freely onto the screen for (b) and (c). However, participants were instructed that their gaze should remain on the screen for the whole time of the trial. For the baseline trials, a fixed presentation duration of 60 seconds was chosen, which roughly corresponds to a reading speed of 200 words per minute (e.g., Rayner et al., 2016; Trauzettel-Klosinski & Dietz, 2012) and thus to the approximate duration of the reading conditions. Each item of the reading conditions was proceeded by a fixation cross (0.5 seconds) that marked the beginning of the first word (grid). Participants were then instructed to fixate each word grid (d) or read every word (e) from top left to bottom right. Regarding the text condition (f), participants were asked to read the newspaper article in a normal manner and at a comfortable pace. There was no time limit for the reading trials, allowing participants to proceeded in a self-paced manner.

Data analysis

The data of the study are available here: https://osf.io/5eysw/.

Preprocessing: Blinks were detected by an algorithm based on pupillometry noise (Hershman et al., 2018) and removed from the data. When more than 10% of data points of a trial were defective, the entire trial was excluded from further analysis. In addition, participants with fewer than three remaining trials per condition were excluded from further analysis. This procedure resulted in the exclusion of one participant and a total of 25 out of 924 trials (2.71%).

As the dependent variable, gaze steps were computed by differencing the raw two-dimensional position data (Stephen & Mirman, 2010). Gaze steps are thus based on consecutive samples of gaze data and not on fixation positions. For instance, the following gaze positions were recorded: $[x_1 = 10, y_1 = 15]$ and $[x_2 = 12, y_2 = 14]$. Here, the gaze step can be calculated as

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{(12 - 10)^2 + (14 - 15)^2} = 2.24.$$

This way, series of position recordings were transformed into series of gaze steps for each trial. Extreme values deviating more than 10 *SD* from the mean of a time series were discarded. Furthermore, fixation durations for trials of the reading conditions were extracted from the data using the Microsaccade Toolbox for R (Engbert et al., 2015). We specified 6 as the minimal number of samples constituting a saccade and used the default velocity factor of 5. Subsequently, both measures were subjected to RQA (Zbilut & Webber, 1992) using the crqa package for R (Coco et al., 2021). Furthermore, SampEn was calculated using a custom script in MATLAB (v2018b). RQA and SampEn were calculated per trial using the parameters described in the following sections.

RQA: In order to run RQA, a delay parameter τ and an embedding parameter D had to be estimated by computing the average mutual information and false nearest neighbor functions. The z-scored data were then subjected to RQA. Following suggestions from Wallot (2017), a threshold parameter T was chosen by an iterative procedure, resulting in a mean RR between 5% and 10% across the whole sample of trials and participants. For gaze step data, the parameters were as follows: $\tau = 7$, $D = 7$, and $T = 0.3$ ($M_{RR} = 7.50$, $SD_{RR} = 5.93$). For fixation duration data, the following parameters were chosen: $\tau = 2$, $D = 3$, and $T = 0.8$ ($M_{RR} = 7.57$, $SD_{RR} = 4.21$). Due to computational limits, RQA for gaze step data was performed in a windowed manner with 10,000 data points at a time in steps of 5,000 data points and then averaged per trial. A tutorial introduction to recurrence quantification analysis is provided by Wallot (2017).

Sample entropy analysis (SampEn). The basis for computing SampEn is calculating the number of matching sequences of some length m and $m + 1$ within a tolerance band defined by a radius r. Both parameters need to be set for analysis (Richman & Moorman, 2000). Here, we determined the length of the template m and the size of the tolerance region r following an approach proposed by Ramdani and colleagues (2009). Regarding our data, we chose $m = 1$ and $r = 3.0$ for gaze step data and $m = 1$ and $r = 3$ for fixation durations. A tutorial introduction to sample entropy analysis is provided by Kuznetsov and colleagues (2013).

Inferential statistics. As can be inferred from hypotheses and design, this study is organized in two parts: a confirmatory part based on gaze step data and an exploratory one based on both gaze step data and fixation durations. Regarding the confirmatory part, we were primarily interested in differences between baseline conditions and reading conditions. Consequently, the respective experimental conditions were subsumed into one overarching factor, with "baseline" and "reading" being the factor levels. However, since the underlying conditions differ from one another, they still were included as a random factor within the multilevel models that were run. For the exploratory part, the individual conditions came into focus, especially the relationship between text grids, shuffled text and normal text. Hence, these conditions were then treated as one fixed factor with three levels in the multilevel models.

The different RQA measures and SampEn, which we obtained for every trial per participant and condition, were subjected to linear mixed-effects models to account for their nested structure (Richter, 2006). The models were set up in RStudio (v1.2.1335) using the lme4 package (v1.1-23) and tested for statistical significance using the lmerTest package (3.1-2). Our model used the following general form:

$$y_{mi} = y_{00} + y_{01} CONT_{mi} + \upsilon_{0i} + \varepsilon_{mi}, \varepsilon \sim N\left(0, \sigma^2\right)$$

Here, y_{00} is the fixed intercept, $y_{01} CONT_{mi}$ is the fixed effect of the contrast of interest, υ_{0i} is the random intercept for participants, and ε_{mi} is the error term. Some of the models also include a random intercept for condition υ_{1i} whenever υ_{1i} contributed significantly to the model.

Results

While the baseline trials were presented with a fixed duration of 60 seconds, the duration of the reading trials depended on individual viewing times. On average, participants spent 82.28 seconds ($SD = 38.75$) on text grids, 65.36 seconds ($SD = 21.38$) on shuffled texts,

		RR		DET		ADL		MDL		SampEn	
Condition	Number of trials	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
(a) Fixation cross	150	3.34	3.27	23.84	18.04	2.13	1.64	68.36	72.72	0.034	0.010
(b) Blank screen	145	4.19	3.26	30.51	17.14	2.69	1.32	81.48	72.24	0.035	0.012
(c) Screen with circles	149	5.01	3.89	37.61	21.98	2.92	1.38	95.92	80.40	0.047	0.014
(d) Text grid	153	10.43	6.56	64.13	24.91	6.19	5.70	204.31	131.20	0.065	0.012
(e) Shuffled text	151	10.78	5.97	69.91	23.12	6.98	6.07	208.41	119.95	0.067	0.009
(f) Normal text	151	11.01	5.43	71.30	20.78	8.40	6.77	215.38	107.41	0.073	0.008
(a–c) Baseline conditions	444	4.18	3.55	30.64	19.96	2.58	1.49	81.89	75.91	0.039	0.013
(d–f) Reading conditions	455	10.74	6.00	68.43	23.17	7.19	6.25	209.34	119.79	0.068	0.010

Table 2. Descriptive statistics for dependent variables based on gaze step data.

	Number of fixations per trial		Fixation duration (ms)		RR		DET		ADL		MDL		SampEn	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
(d) Text grid	253.14	102.09	271.55	37.68	8.15	5.51	20.27	7.01	4.37	1.25	249.14	102.09	0.041	0.007
(e) Shuffled text	251.38	85.53	216.87	19.31	7.20	3.42	17.98	5.02	4.49	0.95	247.38	85.53	0.042	0.009
(f) Normal text	213.48	39.02	205.30	15.63	7.36	3.38	18.13	4.28	4.83	1.40	209.48	39.02	0.041	0.007

Table 3. Descriptive statistics for dependent variables based on fixation duration data.

Characteristic	RR	DET	ADL	MDL	SampEn
RR	—	0.94	0.79	0.95	0.49
DET	0.94	—	0.68	0.92	0.63
ADL	0.79	0.68	—	0.77	0.40
MDL	0.95	0.92	0.77	—	0.50
SampEn	0.49	0.63	0.40	0.50	—

Table 4. Correlation matrix for dependent variables based on gaze step data. *Notes*: Pearson's *r* correlation coefficients. All coefficients are significant at the $p < 0.001$ level.

and 53.23 seconds ($SD = 10.69$) on texts. Descriptive statistics for each dependent variable are provided in Table 2 for gaze step and in Table 3 for fixation duration data. Especially for gaze step data, RQA measures and SampEn showed high intercorrelations (see Table 4), reflecting that they all capture the concept of regularity as was expected. However, these measures are less intercorrelated for fixation durations (see Table 5).

Confirmatory analysis: baseline vs. reading conditions. To test for differences between baseline conditions and reading conditions, linear mixed-effects models were constructed separately for each RQA measure and SampEn. Condition type (baseline vs. reading) was set as categorical fixed effect, and participant and condition were included as random intercepts.

All RQA measures as well as SampEn were affected by condition type (RR: $\chi^2(1) = 20.22$, $^{***}p < 0.001$; DET: $\chi^2(1) = 16.27$, $^{***}p < 0.001$; ADL: $\chi^2(1) = 13.70$, $^{***}p < 0.001$; MDL: $\chi^2(1) = 21.57$, $^{***}p < 0.001$; SampEn: $\chi^2(1) = 13.88$, $^{***}p < 0.001$). All dependent measures distinguished significantly between the two condition types: Compared to reading conditions, baseline conditions exhibit smaller RR, DET, ADL, and MDL, as well as smaller SampEn. Fixed effects for all measures are summarized in Table 6.

The results partially confirmed our hypothesis that reading conditions exhibit higher regularity compared to baseline conditions. Regarding RQA measures, it could be verified that reading conditions lead to higher regularity of eye movement fluctuations as compared to baseline conditions. SampEn results contradicted our prediction if interpreted as a measure of uncertainty. However, if SampEn was interpreted in terms of entropy rate (Porta et al., 2001), it rather captured the complexity of fluctuations, which were potentially related to adaptive cognitive processing.

Exploratory analysis: texts vs. shuffled texts vs. text grids

Gaze step data: In order to determine the extent to which the reading conditions differ from one another, we further set up a linear mixed-effects model for each dependent variable as a function of condition (text vs. shuffled text vs. text grid) as categorical fixed effect. Again, participant was included as random intercept.

While no significant effect of condition could be found for RR and MDL (RR: $\chi^2(2) = 3.50$, $p = 0.174$; MDL: $\chi^2(2) = 3.36$, $p = 0.187$), DET and ADL were affected by condition (DET: $\chi^2(2) = 48.57$,

Measure	Fixation duration	RR	DET	ADL	MDL	SampEn
Fixation duration	—	0.01	0.00	–0.21***	0.33***	–0.07
RR	0.01	—	0.85***	–0.23***	–0.10*	0.03
DET	0.00	0.85***	—	–0.08	–0.29***	0.06
ADL	–0.21***	–0.23***	–0.08	—	–0.41***	0.05
MDL	0.33***	–0.10*	–0.29***	–0.41***	—	–0.05
SampEn	–0.07	0.03	0.06	0.05	–0.05	—

Table 5. Correlation matrix for dependent variables based on fixation duration data. *Notes*: Pearson's *r* correlation coefficients. * $p < 0.05$, *** $p < 0.001$.

Measure		Estimate	*SE*	*df*	*t*	*p*
RR	(Intercept)	10.67	0.92	25.56	11.63	<0.001***
	Baseline	–6.57	0.44	5.12	–14.82	<0.001***
DET	(Intercept)	68.01	4.58	22.72	14.84	<0.001***
	Baseline	–37.73	3.9	5.39	–9.68	<0.001***
ADL	(Intercept)	7.16	0.78	24.59	9.17	<0.001***
	Baseline	–4.61	0.6	5.33	–7.72	<0.001***
MDL	(Intercept)	208.01	18.51	24.6	11.24	<0.001***
	Baseline	–126.84	7.42	5.08	–17.11	<0.001***
SampEn	(Intercept)	0.07	0	8.79	21.73	<0.001***
	Baseline	–0.03	0	5.9	–7.43	<0.001***

Table 6. RQA measures and SampEn for gaze step data: Fixed effects for reading versus baseline conditions. *Notes*: The intercept equals the factor level reading conditions. ***$p < 0.001$.

***$p < 0.001$; ADL: $\chi^2(2) = 35.66$, ***$p < 0.001$). While DET was significantly lower for text grids compared to both normal texts and shuffled texts, it did not differ significantly between normal text and shuffled text. For ADL, a different pattern emerged: It significantly separated normal text from both shuffled text and text grid, but shuffled text and text grid were not distinguishable. Also, SampEn was significantly influenced by reading condition ($\chi^2(2) = 114.54$, ***$p < 0.001$). While SampEn was higher for normal text compared to both other conditions, no differences were found between shuffled text and text grid. See Table 7 for pairwise differences of the fixed factor.

Regarding gaze step data, the RQA results demonstrated that normal text tends to lead to higher regularity of eye movement fluctuations during reading compared to "impoverished" conditions, such as text grid and shuffled text. However, the different RQA measures resulted in distinctive patterns for the conditions, reflecting varying levels of sensitivity. Again, the SampEn results did not follow the pattern as one might expect from a measure of uncertainty or irregularity, but rather complexity.

Fixation durations: Again, linear mixed-effects models for each indicator were computed using condition (normal text, shuffled text, text grid) as categorical fixed effect and participant as random intercept.

While RR only showed a tendency (RR: $\chi^2(2) = 5.71$, $p = 0.057$), DET, ADL, and MDL were affected by condition (DET: $\chi^2(2) = 22.60$, ***$p < 0.001$; ADL: $\chi^2(2) = 13.64$, **$p = 0.001$; MDL: $\chi^2(2) = 34.10$, ***$p < 0.001$). As pairwise tests of fixed effects revealed (see Table 8), normal text exhibited longer ADL but shorter MDL than both other conditions. However, text grid and shuffled text conditions were not significantly different regarding ADL and MDL. DET significantly distinguished text grids from both normal and shuffled texts, with text grids showing higher DET. There was no significant effect for SampEn ($\chi^2(2) = 4.01$, $p = 0.135$).

The results once more indicate that normal reading can be distinguished from related conditions by means of RQA. Opposed to gaze step data, however, the different indicators do not all result in more regularity for normal text. Instead, task-specific patterns emerged. When applied on fixation duration data, SampEn seems noninformative in terms of separating the reading conditions.

Discussion of experiment 1

This study aimed to test the basic assumptions of RTR, namely, that reading of text stimuli leads to higher degrees of regularity compared to baseline conditions where information—and certainly sequentially structured information—was absent. To this end, eye movements were recorded for six conditions, three baseline conditions (fixation cross, blank screen, random circles) and three reading conditions (text grid, shuffled text, normal text). We utilized RQA measures and SampEn, which can be used to capture the strength of regularity from sequential data, and tested these measures on series of gaze steps and fixation durations. Measures and the underlying data type were largely of explorative nature here in order to investigate

Measure	Contrast	Estimate	*SE*	*df*	*t*	*p*
DET	Normal text–shuffled text	1.39	0.98	435	1.42	0.467 n.s.
	Normal text–text grid	6.61	0.98	435	6.78	<0.001***
	Shuffled text–text grid	5.22	0.98	435	5.35	<0.001***
ADL	Normal text–shuffled text	1.43	0.36	435	3.94	<0.001***
	Normal text–text grid	2.16	0.36	435	5.98	<0.001***
	Shuffled text–text grid	0.73	0.36	435	2.03	0.128 n.s.
SampEn	Normal text–shuffled text	0.006	0.001	435	8.61	<0.001***
	Normal text–text grid	0.007	0.001	435	10.81	<0.001***
	Shuffled text–text grid	0.001	0.001	435	2.18	0.089 n.s.

Table 7. RQA measures (DET and ADL) and SampEn for gaze step data: Pairwise comparison of reading conditions. *Notes*: *p*-values were adjusted using the Bonferroni method for three estimates. ***$p < 0.001$, n.s. = not significant.

Measure	Contrast	*Estimate*	*SE*	*df*	*t*	*p*
DET	Normal text—shuffled text	0.22	0.54	442	0.41	1.000 n.s.
	Normal text—text grid	–2.12	0.54	442	–3.94	<0.001***
	Shuffled text—text grid	–2.34	0.54	442	–4.35	<0.001***
ADL	Normal text—shuffled text	0.32	0.12	442	2.67	0.024*
	Normal text—text grid	0.43	0.12	442	3.57	0.001**
	Shuffled text—text grid	0.11	0.12	442	0.91	1.000 n.s.
MDL	Normal text—shuffled text	–38.31	7.57	442	–5.06	<0.001***
	Normal text—text grid	–39.51	7.57	442	–5.22	<0.001***
	Shuffled text—text grid	–1.21	7.57	442	–0.16	1.000 n.s.

Table 8. RQA measures for fixation duration data: Pairwise differences of conditions. *Notes*: *p*-values were adjusted using the Bonferroni method for three estimates. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s. = not significant.

which combination proves most sensitive for future applications of RTR to text reading.

Based on RTR, we predicted lower degrees of regularity for baseline compared to reading conditions. This was tested on gaze step data and largely supported by recurrence measures, with reading conditions exhibiting higher recurrence properties than baseline conditions. For SampEn, we assumed that higher regularity of the reading conditions would be reflected in lower SampEn values. However, the opposite pattern emerged: Reading conditions were more entropic than baseline conditions. Furthermore, we anticipated both text grids and shuffled texts to have lower degrees of regularity compared to normal text. Since the computed regularity measures were not used in this research area before, these assumptions were of an exploratory nature. Support for these predictions was mixed: Normal text showed higher recurrence properties and SampEn values compared to randomized texts and text grids for the gaze step data. For fixation data, however, DET and MDL showed opposite patterns of results (i.e., lower regularity for normal text) while ADL confirmed the expected pattern again. SampEn showed no significant effect at all. Thus, the effects observed for series of fixation durations were rather inconclusive, with recurrence measures showing positive, negative, and null effects, and null effects for entropy measures throughout.

Even though we found supporting evidence for our hypotheses, this support is weakened by the exploratory character of the analysis, as it rested on the post hoc selected combination of measures and data type. Hence, confirmatory studies are needed to strengthen this evidence.

Data type

Regarding the comparison of data type (gaze steps vs. fixation durations), our results clearly favored gaze step data. First, results based on series of gaze steps were generally more sensitive to our manipulations (recurrence and entropy measures yielded significant differences between condition types and among reading conditions), while RR and SampEn did not distinguish between our manipulations when calculated for fixation durations. This might partially be grounded in data size requirements: Gaze step data comprised several

thousand data points per trial, whereas series of fixation durations consisted of fewer than 200 data points.

Second, the direction of effects was more in line with the predictions of RTR. Reading conditions resulted in higher degrees of regularity compared to baseline conditions when the analyses were based on gaze step data, SampEn posing an exception. When based on fixation durations, this was only true for ADL while RR and SampEn yielded null effects, and even the opposite pattern was found for DET and MDL. It might be the case that this is a result of comparatively short trial length. There are startup transients in reading tasks that span over multiple up to several hundred fixations of word reading times, leading to initially higher variability in reading task performance as would be expected for the whole task (Wallot et al., 2013, 2019). Also, different tasks produce somewhat different eye movement dynamics, and parsing such records can sometimes lead to systematically different estimates of fixation durations (Karsh & Breitenbach, 2021).

Finally, gaze step data were more versatile than fixation durations and can be used to compare qualitatively different tasks, some of which might not exhibit fixation- and saccade-like properties such as the baseline conditions that we used here.

Conditions and instructions

The assumptions spelled out in A1 to A4 rested on the idea of a baseline measure for eye movements, meaning absence of external information. While we tried to create three reasonable baseline conditions that were low on what can be thought of as external information, they still provide varying degrees of information to structure gaze activity. While it is probably impossible to talk about eye movements in the absence of external information in the strict sense, it would be helpful to have a general metric on information that could be applied in order to quantify the distance between the baseline and reading conditions in this regard.

Also, the chosen reading conditions offered only a first and limited insight in applying recurrence and entropy measures to the reading process. The conditions chosen did not resemble a continuous range from "information-free" contexts toward a full, naturalistic text presentation. Such an investigation would surely be interesting when focusing on variants of text-like conditions in order to clarify what different text features contribute to the reading process. However, with regard to the feasibility of this study, we had to restrict the set of conditions to some relevant contrasts for the central research question asked here, since our goal was not yet to map out the influence of different text properties on RTR, but first and foremost to establish an understanding of regularity in contexts with minimal external information compared to the processing of text-like variations and actual texts.

Furthermore, task instructions between the experimental conditions varied so that participants behaved most properly within each condition. However, this might limit the conclusions that can be drawn from the experiment, as participants' behavior was now a function of stimuli and instruction together. The decision to use different instructions was motivated by the fact that participants can handle stimuli quite differently when not explicitly instructed. During the pilot phase of the experiment, participants were more comfortable letting their gaze wander or looking at a different part of the screen instead of staring at the displayed fixation cross for the entire 60 seconds of a trial. Similarly, participants did not necessarily engage in reading-like behavior when text grids or random text was presented, but rather let their gaze wander or even jumped back and forth in an attempt to puzzle together a meaningful text. While these spontaneous interaction patterns with different stimuli were quite fascinating, they were not pertinent to tackle the underlying research question. Still, in order to address the question of how instructions might have contributed to the observed pattern of results, we conducted a second study with a uniform instruction across conditions.

Experiment 2

In order to address the points of varying instructions and a limited set of conditions as discussed above, we carried out an additional study. A more general but uniform instruction was used to distinguish effects driven by instructions and effects due to linguistic information conveyed by the different stimuli. Specifically, participants were told to look at the contents presented on the screen, irrespective of the particular stimulus type. Furthermore, a more differentiated set of conditions reflecting a more graduated buildup of linguistic information was chosen for this second study. At the same time, this posed a chance to corroborate the findings of the previous study and to further explore the sensitivity of measures of RTR.

Hypotheses

This second study further investigated the differences captured by measures of regularity for conditions that reflect more graduated levels of external linguistic information available in a stimulus (see Figure 2). Based on the concept of RTR and the previous findings of Experiment 1, we expected strongest regularity for normal text reading. Based on our reasoning from the

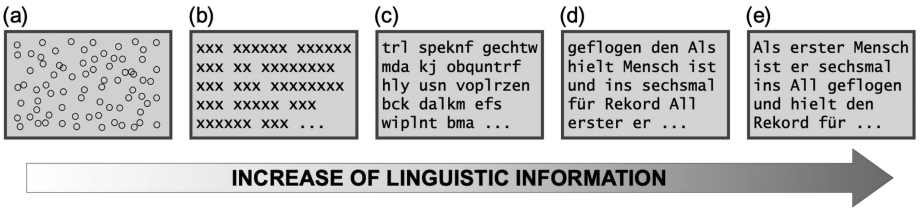


Figure 2. Schematic examples for the experimental conditions. Experimental conditions consisted of (a) circles, (b) text grid, (c) pseudo text, (d) randomized text, and (e) normal text.

previous study, we expected to find more regularity in those conditions more similar to normal text. However, we have to cautiously qualify this hypothesis. Not providing participants with specific instructions of what to do might lead to different patterns of behavior. For example, eye movements differ greatly if participants read a text for comprehension, search for typos, or count the number of words in a text. While we were intuitively confident that participants would engage in normal reading behavior when presented with an actual text (this should be what skilled readers are naturally inclined to do), it was less clear how they would act in the less self-instructing conditions.

Again, regularity was operationalized by means of RQA measures (i.e., RR, DET, ADL, and MDL) as well as SampEn that were computed based on series of gaze steps. This combination of measures and data type proved most suitable to capture the different degrees of linguistic information reflected in eye movement data in Experiment 1.

Method

Participants

Twenty-seven German natives with normal or corrected-to-normal vision participated in the study. They did not take part in the previous experiment and had no neuropsychological disabilities. Participants were compensated for their time with 7€ per 30 minutes and received 14€ on average. One participant had to be excluded due to problems during the calibration procedure. Three more participants dropped out of analysis due to excessive blinking artifacts in the recorded data. Thus, the final sample consisted of 23 participants (13 female) with a mean age of 26.43 years ($SD = 4.97$). See Appendix A for further information about the participants. Written informed consent was obtained from all participants prior to the experiment. As for the previous study, the method was approved by the Ethics Council of the Max Planck Society and followed the ethical principles of the Declaration of Helsinki.

Stimuli

All in all, Experiment 2 comprised five conditions: (a) circles, (b) text grid, (c) pseudo text, (d) shuffled text, and (e) normal text. Except for the pseudo text condition, all other conditions were part of Experiment 1 (see above for a detailed description of stimulus selection and generation). The pseudo text condition was included in order to decrease the leap between text grids and shuffled texts. While text grids preserved the general layout of a text (all letters replaced by "x" but spatial organization through spaces and lines kept intact), shuffled texts already contained semantic information on the word and topic level (randomized word order of actual newspaper articles). For the pseudo text condition, words of a text were replaced by random letter strings that do not constitute any German words and are unpronounceable for German natives.

Procedure

The study was carried out with the same spatial and technical setup as described above for Experiment 1. It took participants about 50 minutes to complete the experiment, including a short break halfway through the experiment. Again, participants were randomly distributed to one of two stimulus lists: Actual newspaper articles assigned to List A served as text base for conditions (b) to (e) in List B and vice versa. The experiment comprised 7 trials per condition, resulting in 35 trials in total. All trials were presented in a fully randomized order.

Participants were instructed to look at the content presented on the screen and that their gaze should remain on the screen during the entire trial. Since participants were intentionally not instructed to read in any of the conditions, there was no fixation cross preceding any of the trials. Furthermore, trial duration was set uniformly to 40 seconds. This time interval

was deliberately chosen to be shorter than the average reading times obtained from Experiment 1 in order to prevent fast-reading participants from finishing before the end of the trial.

Data analysis

The data of the study are available here: https://osf.io/5eysw/.

Preprocessing: All steps regarding preprocessing were kept the same as in Experiment 1, so that a certain comparability of data and results was ensured. Due to blinks and artifacts that were detected based on the pupillometry noise algorithm (Hershman et al., 2018), data of three participants were discarded, and a total of 24 out of the remaining 805 trials (2.98%) was excluded from further data analysis. In a trial-by-trial manner, gaze steps were calculated (cf. Stephen & Mirman, 2010), and extreme values that differed more than 10 *SD* from the mean were removed. Since fixation durations turned out to be less well suited to capture the eye movement dynamics of interest in Experiment 1, these were not extracted for Experiment 2.

RQA and SampEn: Time series of gaze steps were subjected to RQA and SampEn analysis using the same resources as for Experiment 1, that is, the crqa package for R (Coco et al., 2021) and a custom-script for MATLAB to compute SampEn. Again, a windowed RQA was computed with a window size of 10,000 data points and a window step of 5,000 data points. Afterward, RQA measures were averaged per trial. Based on an iterative procedure, the following parameters were specified: a delay parameter $\tau = 2$, an embedding parameter $D = 4$, and a threshold parameter $T = 0.5$. These parameters resulted in a mean RR of 7.30% ($SD_{\mathrm{RR}} = 8.25$) for the whole sample. SampEn analysis was carried out with a template length $m = 1$ and a size of the tolerance region $r = 3.0$ (cf. Ramdani et al., 2009).

Inferential statistics: As described above, this second study investigated differences in regularity measures between five experimental conditions. Regularity was operationalized by means of the RQA measures RR, DET, ADL, and MDL, as well as SampEn. Each of these dependent variables was subjected to linear mixed-effects models using the R packages lme4 (v1.1-23) and lmerTest (3.1-2). Within the multilevel models, condition was defined as fixed factor with five levels, and a random intercept for participants was included, according to the following general form:

$$y_{mi} = y_{00} + y_{01} COND_{mi} + \upsilon_{0i} + \varepsilon_{mi}, \varepsilon \sim N\left(0, \sigma^2\right)$$

Here, y_{00} is the fixed intercept, $y_{01} COND_{mi}$ is the fixed effect for condition, υ_{0i} is the random intercept for participants, and ε_{mi} is the error term.

Results

Table 9 provides the descriptive statistics for all dependent measures. Condition affected all regularity measures but MDL (RR: $\chi^2(4) = 224.53$, $^{***}p < 0.001$; DET: $\chi^2(4) = 283.00$, $^{***}p < 0.001$; ADL: $\chi^2(4) = 54.47$, $^{***}p < 0.001$; SampEn: $\chi^2(4) = 289.49$, $^{***}p < 0.001$; MDL: $\chi^2(4) = 6.00$, $p = 0.199$). For RR and ADL, values gradually increased the more linguistic information became available. Apart from two contrasts (circles vs. text grid and text grid vs. pseudo text), all other pairwise comparisons were significant. While descriptive results for ADL revealed a similar pattern, only the contrasts of normal text compared to pseudo text, text grid and circles, and random text compared to text grid and circles reached significance. SampEn did not differentiate pseudo text from text grid and circles, but it still exhibited the expected pattern of results for all other contrasts. See Table 10 for pairwise differences of the fixed factor. These findings supported the hypothesis that normal text exhibits more regularity than the other conditions. Furthermore, results mostly support the assumption that increasing availability of external linguistic information leads to increased regularity that can be meaningfully depicted by means of recurrence and entropy measures.

As shown in Table 11, intercorrelations of regularity measures were overall high with the exception of SampEn and MDL, which showed rather moderate correlations strengths. This basically replicated findings from Experiment 1 suggesting that the utilized measures indeed capture the regularity concept well and to a similar degree.

Discussion of experiment 2

This second study provided additional evidence for how measures of regularity can reliably capture varying degrees of linguistic information conveyed by visual stimuli in time-series data. Five experimental conditions were chosen, with arbitrary layouts of circles providing no linguistic context at all, and, opposed to that, short newspaper articles incorporating the maximum of linguistic information represented the extrema. Three conditions in between, text grids, pseudo texts, and texts with randomized word order, comprised increasing levels thereof. Again, recurrence and entropy measures were used to capture the strength of regularity based on series of gaze steps.

We hypothesized that regularity measures should be highest for normal text and lower for the other conditions. This prediction was borne out by the observed results. Furthermore, we more cautiously presumed that increasing linguistic information could be reflected by increasing regularity measures. Also, this assumption was mostly supported by the results. Since

		RR		DET		ADL		MDL		SampEn	
Condition	Number of trials	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Normal text	160	11.98	10.58	26.35	19.98	3.841	1.535	41.73	57.72	0.064	0.011
Shuffled text	157	9.03	8.78	20.37	16.46	3.461	1.271	32.69	49.19	0.059	0.012
Pseudo text	154	5.83	7.07	13.57	13.73	3.195	2.002	40.45	184.20	0.051	0.013
Text grid	152	5.10	5.35	12.41	11.81	2.956	0.648	19.74	14.26	0.052	0.013
Circles	158	4.42	5.69	10.70	11.45	2.983	1.627	26.42	103.88	0.049	0.011

Table 9. Descriptive statistics for dependent variables

Measure	Contrast	Estimate	*SE*	*df*	*t*	*p*
RR	Normal text–shuffled text	3.01	0.55	762	5.46	<0.001***
	Normal text–pseudo text	6.03	0.56	762	10.849	<0.001***
	Normal text–text grid	6.78	0.56	762	12.155	<0.001***
	Normal text–circles	7.66	0.55	762	13.889	<0.001***
	Shuffled text–pseudo text	3.01	0.56	762	5.40	<0.001***
	Shuffled text–text grid	3.76	0.56	762	6.72	<0.001***
	Shuffled text–circles	4.65	0.55	762	8.38	<0.001***
	Pseudo text–text grid	0.75	0.56	762	1.33	1.000 n.s.
	Pseudo text–circles	1.63	0.56	762	2.93	0.035*
	Text grid–circles	0.89	0.56	762	1.58	1.000 n.s.
DET	Normal text–shuffled text	6.10	0.99	762	6.13	<0.001***
	Normal text–pseudo text	12.52	1.00	762	12.522	<0.001***
	Normal text–text grid	13.70	1.00	762	13.659	<0.001***
	Normal text–circles	15.87	0.99	762	15.984	<0.001***
	Shuffled text–pseudo text	6.42	1.00	762	6.39	<0.001***
	Shuffled text–text grid	7.60	1.01	762	7.55	<0.001***
	Shuffled text–circles	9.77	1.00	762	9.79	<0.001***
	Pseudo text–text grid	1.18	1.01	762	1.17	1.000 n.s.
	Pseudo text–circles	3.35	1.00	762	3.34	0.009**
	Text grid–circles	2.16	1.01	762	2.15	0.317 n.s.
ADL	Normal text–shuffled text	0.39	0.14	762	2.77	0.057 n.s.
	Normal text–pseudo text	0.63	0.14	762	4.53	<0.001***
	Normal text–text grid	0.87	0.14	762	6.24	<0.001***
	Normal text–circles	0.87	0.14	762	6.28	<0.001***
	Shuffled text–pseudo text	0.25	0.14	762	1.76	0.787 n.s.
	Shuffled text–text grid	0.49	0.14	762	3.48	0.005**
	Shuffled text–circles	0.49	0.14	762	3.48	0.005**
	Pseudo text–text grid	0.24	0.14	762	1.72	0.868 n.s.
	Pseudo text–circles	0.24	0.14	762	1.70	0.892 n.s.
	Text grid–circles	–0.00	0.14	762	–0.03	1.000 n.s.
SampEn	Normal text–shuffled text	0.005	0.001	762	5.50	<0.001***
	Normal text–pseudo text	0.013	0.001	762	13.24	<0.001***
	Normal text–text grid	0.012	0.001	762	12.94	<0.001***
	Normal text–circles	0.015	0.001	762	16.04	<0.001***
	Shuffled text–pseudo text	0.007	0.001	762	7.73	<0.001***
	Shuffled text–text grid	0.007	0.001	762	7.46	<0.001***
	Shuffled text–circles	0.010	0.001	762	10.48	<0.001***
	Pseudo text–text grid	0.000	0.001	762	–0.25	1.000 n.s.
	Pseudo text–circles	0.003	0.001	762	2.68	0.075 n.s.
	Text grid–circles	0.003	0.001	762	2.92	0.036*

Table 10. RQA measures (RR, DET, and ADL) and SampEn: Pairwise comparisons. *Notes*: *p*-values were adjusted using the Bonferroni method for 10 estimates. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s. = not significant.

Measure	RR	DET	ADL	MDL	SampEn
RR	—	0.99***	0.83***	0.50***	0.50***
DET	0.99***	—	0.81***	0.46***	0.56***
ADL	0.83***	0.81***	—	0.86***	0.24***
MDL	0.50***	0.46***	0.86***	—	0.00 n.s.
SampEn	0.50***	0.56***	0.24***	0.00 n.s.	—

Table 11. Correlation matrix for dependent variables. *Notes*: Pearson's *r* correlation coefficients. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s. = not significant.

these results were observed when task instructions were kept constant across conditions, we can be confident in the validity of the findings of Experiment 1. At the same time, however, the uniform instructions impede a further interpretation of significant effects (or the lack thereof) for some of the conditions with intermediate linguistic information (i.e., with regard to differences between shuffled texts, pseudo texts, and text grids).

General discussion and outlook

The central aim of the present article was to test a fundamental assumption of RTR. That is, with enhancing degrees of external (linguistic) information, the regularity of dynamical measures that reflect processing during reading increases. To prove this, we used measures that capture the regularity enclosed in time series, here specifically measures of recurrence and entropy. These measures were applied to eye movements that we recorded for contexts in which linguistic information was absent, increasingly text-like conditions, and the presentation of actual texts. Findings across two experiments showed that regularity measures distinguished successfully between text reading and conditions with varying degrees of linguistic information. However, some specific patterns of results emerged for the different regularity measures that need to be further discussed. In particular, SampEn did not behave in a way that warrants a plain interpretation in terms of regularity. Furthermore, we would like to discuss the limitations of the studies reported here and provide an outlook for future research.

Measures

Conceptually, recurrence and entropy measures imply a fairly straightforward interpretation: Higher regularity in a time series of eye movements should be reflected in higher values for RQA measures and lower values for SampEn. And indeed, the first part of this notion was supported by our results: Recurrence measures consistently indicated higher regularity for reading conditions compared to baseline conditions and, for gaze step data, also higher regularity the more similar stimuli were to normal texts. However, results for SampEn opposed this tenet. While SampEn did prove to be a sensitive measure for regularity, its effects seemed to contradict the concept of RTR.

A possible explanation for this might be that SampEn is, strictly speaking, not a classical entropy measure. As pointed out in the Introduction, the calculation of SampEn is based on how well smaller templates in a time series extend to larger ones. Hence, it might be more similar to measures of entropy rate (Porta et al., 2001) than to entropy measures per se. As entropy rate captures complexity of data (i.e., the presence of multiple systematic patterns in a time series), it rather captures complexity of a signal and indexes adaptive cognitive processing but not irregularity.

What does this imply? One of the exploratory aims of the current study was to use different potentially suitable measures to capture RTR and investigate which of these prove to be sensitive. While SampEn did turn out to capture the dynamics of interest, the direction of effects is not easily reconcilable with the notion of RTR. If SampEn would indeed be interpreted as a complexity measure, it might capture an aspect of skilled reading that is not (yet) incorporated into the concept of RTR, namely, adaptive flexibility. As outlined above, RTR focusses on the stability of reading behavior over time that is expected to arise from skilled reading. However, skill behavior also has an adaptive component that is not reflected within stability, that is, skill execution of behavior also entails quick and successful adaption to changes in the situation (Riley & Turvey, 2002; Ward et al., 2018).

Interpreted this way, SampEn as a complexity measure might rather capture this adaptability facet of skill. Consequently, skilled reading would be marked by high stability of the process but, at the same time, by high adaptive flexibility. This notion would also be in line with findings that multifractal measures that capture complexity of behavior (e.g., Ihlen & Vereijken, 2010; Kelty-Stephen & Wallot, 2017) are also increased in high-skilled readers (Wallot et al., 2014). However, this train of thought warrants a theoretical expansion of the RTR concept that has yet to be conceptualized.

Limitations

The conclusions that can be drawn from the current studies are limited by several factors. First of all, the assumptions spelled out in A1 to A4 rest on the idea

of a baseline measure for eye movements as such, that is, the absence of external information. While we tried to create three reasonable baseline conditions low on what can be thought of as external information, they do still provide varying degrees of information to structure gaze activity. While it is probably impossible to talk about eye movements in the absence of external information in a strict sense, it would be helpful to have a general metric on information that could be used to quantify the distance of the baseline conditions and reading conditions in this regard. Also, while we find supporting evidence for our hypotheses, this support is weakened by the exploratory character of the analysis, as it rests on the post hoc selected combination of measures and data type. Hence, confirmatory studies are needed to strengthen this evidence.

Here, it also has to be mentioned that the current approach on evaluating regularity metrics rests on individual evaluations in separate univariate analyses. While this serves our goal of identifying which of these metrics are suitable and sensitive operationalizations of RTR, a multivariate combination of these measures might yield further insights or even better separability of conditions.

Furthermore, the results based on gaze step and fixation durations of the first experiment are not fully comparable. Some of the metrics employed here gain in reliability with increasing length of a time series. Accordingly, results based on gaze steps might merely be more sensitive to the experimental manipulations by virtue of greater time series length compared to fixation-based results.

Finally, RTR was formulated for the application to reading tasks (Wallot 2014, 2016), especially to connected text reading. However, text stimuli of the current study consisted of only relatively short newspaper articles that tend to work differently than longer connected texts (Wallot et al., 2013, 2019). Accordingly, future studies need to validate the current findings on longer text stimuli.

Outlook

In the current studies, we introduced RTR as a means to capture the process of connected text reading. Our results support that RTR adequately captures the difference between nonreading and reading conditions, as well as show evidence for the assumption that sequential information inherent in text reading leads to stronger regularity of reading process measures. Furthermore, our results suggest that recurrence measures and SampEn are well-suited measures to capture RTR. Moreover, when using eye movements, gaze step data seem to be the better basis for such analyses compared to series of fixation durations.

However, reading ultimately pursues the goal of gaining information, that is, comprehending a text. Thus, the motivation for RTR originates in text comprehension research and the questions of whether and how comprehension can be predicted by means of process measures of reading across tasks (Teng et al., 2016) and languages (Frost, 2012). On the one hand, various measures of the reading process such as word or sentence reading times, fixation durations, or the number of regressive eye movements have been shown to vary with local or global text difficulty (e.g., Just & Carpenter, 1980; Rayner et al., 2006). Using such measures to predict comprehension, on the other hand, has been far from trivial and did not always succeed (LeVasseur et al., 2006, 2008).

Some studies that utilized regularity metrics had some success in predicting comprehension from reading times and eye movements (Wallot et al., 2014, 2015). The current article was based on this work. But also other recent studies have successfully predicted comprehension using the notion of coupling between text features and perceptual-cognitive processing. For instance, Mills and colleagues (2017) showed that reading times and cognitive coupling, operationalized as regression of reading times and text complexity, were positive predictors of participants' reading comprehension. Moreover, they demonstrated that decoupling measured in instances of mind-wandering resulted in worse text comprehension. Moreover, Southwell and colleagues (2020) showed that comprehension scores can be successfully predicted from reading times and classical eye movement measures. However, it remains unclear why the same measures yielded null effects in other studies (Wallot et al., 2015) or related reading speed components during self-paced reading (LeVasseur et al., 2006, 2008; Wallot et al., 2014). Potentially, this might be traced back to differences in modeling and sample size, but also to how comprehension was assessed, and the parameter settings applied to define reading times or extract fixations.

Conceptually, we do see a potential advantage for RTR-based measures because they do not depend on defining text properties whose effects might not be independent of task and language. However, whether RTR offers better metrics to predict reading comprehension from process data is an empirical question that will have to be addressed in future studies, investigating the relation between the reading process and comprehension, directly comparing the different successful approaches on the same data sets but also across important variations such as different types of reading tasks and writing systems.

Keywords: reading time regularity, information processing, recurrence quantification analysis, sample entropy analysis, text reading

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) by grants to Sebastian Wallot (project numbers 397523278 and 442405852).

https://orcid.org/0000-0002-2434-4516 (MT)
https://orcid.org/0000-0002-3626-3940 (SW)

Commercial relationships: none.
Corresponding author: Sebastian Wallot.
Email: sebastian.wallot@leuphana.de.
Address: Leuphana University of Lüneburg, Institute of Psychology, Lüneburg, Germany.

References

Booth, C. R., Brown, H. L., Eason, E. G., Wallot, S., & Kelty-Stephen, D. G. (2018). Expectations on hierarchical scales of discourse: Multifractality predicts both short- and long-range effects of violating gender expectations in text reading. *Discourse Processes, 55*(1), 12–30.

Coco, M. I., Mønster, D., Leonardi, G., Dale, R., & Wallot, S. (2021). Unidimensional and multidimensional methods for recurrence quantification analysis with CRQA. *R Journal, 13*(1), 145–163, doi:10.32614/RJ-2021-062.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*(4), 777–813.

Engbert, R., Sinn, P., Mergenthaler, K., & Trukenbrod, H. (2015). Microsaccade Toolbox for R. *Potsdam Mind Research Repository,* https://t1p.de/ochq.

Faber, M., Krasich, K., Bixler, R. E., Brockmole, J. R., & D'Mello, S. K. (2020). The eye–mind wandering link: Identifying gaze indices of mind wandering across tasks. *Journal of Experimental Psychology: Human Perception and Performance, 46*(10), 1201–1221.

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*(5), 263–279.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, 36*(2), 193–202.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*(3), 518–565.

Heister, J., Würzner, K. M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., . . . Kliegl, R. (2011). dlexDB—eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau, 62*(1), 10–20.

Hershman, R., Henik, A., & Cohen, N. (2018). A novel blink detection method based on pupillometry noise. *Behavior Research Methods, 50*(1), 107–114.

Holden, J. G., & Van Orden, G. (2002). Reading. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 951–955). Cambridge, MA: MIT Press.

Ihlen, E. A., & Vereijken, B. (2010). Interaction-dominant dynamics in human cognition: Beyond $1/f\alpha$ fluctuation. *Journal of Experimental Psychology: General, 139*(3), 436.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.

Karsh, R., & Breitenbach, F. W. (2021). Looking at looking: The amorphous fixation measure. In *Eye movements and psychological functions* (pp. 53–64). Abingdon: Routledge.

Kelty-Stephen, D. G., & Wallot, S. (2017). Multifractality versus (mono-) fractality as evidence of nonlinear interactions across timescales: Disentangling the belief in nonlinearity from the diagnosis of nonlinearity in empirical data. *Ecological Psychology, 29*(4), 259–299.

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5*(3), 257–274.

Kuznetsov, N., Bonnette, S., & Riley, M.A. (2013). Nonlinear time series methods for analyzing behavioural sequences. In K. Davids, R. Hristovski, D. Araújo, N. Balague Serre, C. Button, & P. Passos (Eds.), *Complex systems in sport* (pp. 111–130). Abingdon: Routledge.

LeVasseur, V. M., Macaruso, P., Palumbo, L. C., & Shankweiler, D. (2006). Syntactically cued text facilitates oral reading fluency in developing readers. *Applied Psycholinguistics, 27*(3), 423–445.

LeVasseur, V. M., Macaruso, P., & Shankweiler, D. (2008). Promotion gains in reading fluency: A comparison of three approaches. *Reading and Writing, 21*(3), 205–230.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports, 438*(5–6), 237–329.

McNerney, M. W., Goodwin, K. A., & Radvansky, G. A. (2011). A novel study: A situation model analysis of reading times. *Discourse Processes, 48*(7), 453–474.

Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General, 146*(6), 872–883.

O'Brien, B. A., & Wallot, S. (2016). Silent reading fluency and comprehension in bilingual children. *Frontiers in Psychology, 7*, 1265.

O'Brien, B. A., Wallot, S., Haussmann, A., & Kloos, H. (2014). Using complexity metrics to assess silent reading fluency: A cross-sectional study comparing oral and silent reading. *Scientific Studies of Reading, 18*(4), 235–254.

Porta, A., Guzzetti, S., Montano, N., Furlan, R., Pagani, M., Malliani, A., … Cerutti, S. (2001). Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *IEEE Transactions on Biomedical Engineering, 48*(11), 1282–1291.

Ramdani, S., Seigle, B., Lagarde, J., Bouchara, F., & Bernard, P. L. (2009). On the use of sample entropy to analyze human postural sway data. *Medical Engineering & Physics, 31*(8), 1023–1031.

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*(1), 241–255.

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 787–799.

Rayner, K., Schotter, E. R., Masson, M. E., Potter, M. C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest, 17*(1), 4–34.

Reichle, E. D., Liversedge, S. P., Drieghe, D., Blythe, H. I., Joseph, H. S., White, S. J., … Rayner, K. (2013). Using EZ Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review, 33*(2), 110–149.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*(1), 1–21.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology–Heart and Circulatory Physiology, 278*(6), H2039–H2049.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*(3), 221–250.

Riley, M. A., & Turvey, M. T. (2002). Variability and determinism in motor behavior. *Journal of Motor Behavior, 34*(2), 99–125.

Schad, D. J., Nuthmann, A., & Engbert, R. (2012). Your mind wanders weakly, your mind wanders deeply: Objective measures reveal mindless reading at different levels. *Cognition, 125*(2), 179–194.

Southwell, R., Gregg, J., Bixler, R., & D'Mello, S. K. (2020). What eye movements reveal about later comprehension of long connected texts. *Cognitive Science, 44*(10), e12905.

Stephen, D. G., & Mirman, D. (2010). Interactions dominate the dynamics of visual cognition. *Cognition, 115*(1), 154–169.

Teng, D. W., Wallot, S., & Kelty-Stephen, D. G. (2016). Single-word recognition need not depend on single-word features: Narrative coherence counteracts effects of single-word features that lexical decision emphasizes. *Journal of Psycholinguistic Research, 45*(6), 1451–1472.

Trauzettel-Klosinski, S., & Dietz, K. (2012). Standardized assessment of reading performance: The new international reading speed texts IReST. *Investigative Ophthalmology & Visual Science, 53*(9), 5452–5461.

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual differences in eye-movements during reading: Working memory and speed-of-processing effects. *Journal of Eye Movement Research, 5*(1), 1–16.

Velan, H., & Frost, R. (2007). Cambridge University versus Hebrew University: The impact of letter transposition on reading English and Hebrew. *Psychonomic Bulletin & Review, 14*(5), 913–918.

Wallot, S. (2014). From "cracking the orthographic code" to "playing with language": Toward a usage-based foundation of the reading process. *Frontiers in Psychology, 5,* 891.

Wallot, S. (2016). Understanding reading as a form of language-use: A language game hypothesis. *New Ideas in Psychology, 42*, 21–28.

Wallot, S. (2017). Recurrence quantification analysis of processes and products of discourse: A tutorial in R. *Discourse Processes, 54*(5–6), 382–405.

Wallot, S., Hollis, G., & van Rooij, M. (2013). Connected text reading and differences in text

reading fluency in adult readers. *PLoS One, 8*(8), e71914.

Wallot, S., Lee, J. T., & Kelty-Stephen, D. G. (2019). Switching between reading tasks leads to phase-transitions in reading times in L1 and L2 readers. *PLoS One, 14*(2), e0211502.

Wallot, S., O'Brien, B., Coey, C. A., & Kelty-Stephen, D. (2015). Power-law fluctuations in eye movements predict text comprehension during connected text reading. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, . . . P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2583–2588). Austin, TX: Cognitive Science Society.

Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, A., & Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1745–1765.

Ward, P., Gore, J., Hutton, R., Conway, G. E., & Hoffman, R. R. (2018). Adaptive skill as the conditio sine qua non of expertise. *Journal of Applied Research in Memory and Cognition, 7*(1), 35–50.

Whitney, C., & Cornelissen, P. (2005) Letter-position encoding and dyslexia. *Journal of Research in Reading, 28*(3), 274–301.

Wijnants, M. L., Hasselman, F., Cox, R. F. A., Bosman, A. M. T., & Van Orden, G. (2012). An interaction-dominant perspective on reading fluency and dyslexia. *Annals of Dyslexia, 62*(2), 100–119.

Yarbus, A. L. (2013). *Eye movements and vision*. New York: Springer. (Original work published 1967)

Zbilut, J. P., & Webber, C. L., Jr. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A, 171*(3–4), 199–203.

Ziegler, J. C., Perry, C., & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology, 12*(3), 413–430.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*(2), 386–397.

Appendix A: Participant information

Experiment	Participant	List	Sex	Age (years)	Languages	Education	Vision problems	Vision aid	Reading (hours per week)
Experiment 1	1	A	male	24	3	2	no	none	45
	2	B	female	21	3	1	yes	contacts	20
	3	B	female	21	3	1	no	none	10
	4	A	female	29	4	2	no	glasses	10
	6	B	female	28	3	2	no	none	24
	7	A	male	37	4	1	yes	glasses	40
	9	A	female	25	5	2	no	none	4
	10	B	male	21	3	1	no	none	6
	11	A	female	23	3	1	no	none	14
	15	A	female	28	2	1	yes	contacts	12
	16	A	female	57	2	3	no	glasses	15
	17	A	female	25	3	2	no	glasses	35
	18	B	male	28	3	2	no	none	18
	19	B	male	23	2	2	no	none	10
	20	B	female	34	2	1	no	none	30
	21	A	male	21	2	2	no	none	20
	23	A	male	51	2	3	yes	glasses	15
	24	B	female	24	4	1	no	none	3
	25	B	female	21	3	1	no	none	21
	26	B	female	22	3	1	no	none	14
	27	B	female	24	4	1	no	none	10
	28	B	female	21	4	1	yes	contacts	12
Experiment 2	2	B	male	29	4	2	yes	contacts	30
	3	A	male	24	4	2	no	none	8
	5	A	female	24	2	1	yes	glasses	60
	6	B	female	20	5	1	no	none	12
	7	A	female	21	3	1	no	none	14
	8	B	female	23	4	1	no	none	18
	10	B	female	29	4	2	yes	contacts	14
	11	A	female	29	3	2	yes	glasses	14
	12	B	male	22	3	1	no	none	8
	13	A	female	27	4	1	yes	contacts	8
	14	B	male	31	4	2	yes	glasses	20
	15	A	female	23	3	2	no	none	10
	16	B	female	23	2	1	yes	glasses	25
	17	A	female	38	3	3	no	none	14
	18	B	male	24	2	2	no	none	12
	19	A	male	24	4	1	yes	glasses	5
	20	B	female	22	5	1	no	none	21
	22	B	female	31	4	4	no	none	10
	24	B	female	30	3	2	no	none	15
	26	B	male	25	4	1	yes	contacts	20
	27	A	male	39	3	3	no	none	30
	28	B	male	24	2	1	yes	glasses	14
	29	A	male	26	4	2	yes	glasses	14

Notes: All sociodemographic variables were collected through a short survey after completion of the experiment. Reported values were self-indications by the participants. “List” refers to the assigned list of texts; “Languages” refers to the number of languages participants could fluently speak and read; “Education” indicates participants’ highest education level (1: university entrance degree, 2: bachelor’s degree or equivalent, 3: master’s degree or equivalent, 4: others).

Appendix B: Text characteristics

List	Title	Section	Date of publication	Words	Sentences	Clauses	Syllables	Graphemes	Words per sentence	Words per clause	Syllables per word	Graphemes per word	Type frequency		Log type frequency		Annotated type frequency		Log annotated type frequency	
													M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
A	Berufstätigkeit: Mütter arbeiten länger	Economics	Jan. 23, 2018	180	10	16	374	1,039	18.00	11.25	2.08	5.77	31,3759.03	697,674.17	4.18	1.53	285,918.10	645,816.57	4.06	1.54
	Bestseller: Überraschungs-erfolg für Lundes „Bienen“	Feuilleton	Jan. 4, 2018	195	10	14	375	1,109	19.50	13.93	1.92	5.69	492,106.81	949,559.24	3.98	1.96	439,482.36	877,877.76	3.87	1.96
	Geldautomaten: Schäden steigen auf 2,2 Millionen Euro	Finances	Jan. 15, 2018	194	11	17	447	1,192	17.64	11.41	2.30	6.14	304,975.16	642,842.97	3.79	1.92	266,763.17	569,289.14	3.66	1.92
	Soziales: Gegen Pflichtbesuche in KZ-Gedenkstätten	Politics	Jan. 9, 2018	177	11	22	406	1,186	16.09	8.05	2.29	6.70	448,690.65	888,615.04	4.03	1.85	373,845.26	764,229.37	3.81	1.93
	Gesundheit: TV-Köche missachten ständig die Hygiene	Science	Jan. 22, 2018	157	9	19	343	1,008	17.44	8.26	2.18	6.42	349,691.41	747,197.27	3.80	1.89	289,416.01	637,628.42	3.72	1.85
	Mädchen sind „mutig“? Falsch! Eltern empören sich über eine skurrile Schulaufgabe	Society	Jan. 27, 2018	201	15	29	397	1,174	13.40	6.93	1.98	5.84	359,319.15	720,286.47	4.27	1.54	298,636.54	625,302.44	4.11	1.57
	Fußball: Goretzkas Wechsel zum FC Bayern perfekt	Sports	Jan. 20, 2018	189	12	22	389	1,054	15.75	8.59	2.06	5.58	373,833.61	773,071.25	3.93	2.03	285,300.86	632,720.09	3.74	2.02
	Overall List A			184.71	11.14	19.86	390.14	1,108.86	16.83	9.77	2.12	6.02	377,482.26	69,092.81	4.00	0.18	319,908.90	62,937.58	3.85	0.17
B	Niki-Verkauf: Bundesregierung pocht weiter auf Wettbewerb	Economics	Jan. 3, 2018	197	14	20	443	1,234	14.07	9.85	2.25	6.26	447,430.25	877,364.55	3.98	1.96	396,870.81	796,134.97	3.91	1.94
	Berliner Museums-insel: James-Simon-Galerie wieder im Zeitplan	Feuilleton	Jan. 12, 2018	162	10	13	369	1,016	16.20	12.46	2.28	6.27	421,329.65	844,266.78	3.87	1.93	375,649.46	782,299.27	3.75	1.92
	Übernahmen: Milliardärsfamilie sucht Geldgeber	Finances	Jan. 3, 2018	197	12	23	435	1,196	16.42	8.57	2.21	6.07	412,412.61	882,916.58	3.68	2.05	368,480.32	803,229.19	3.52	2.06
	Russland: Beschützer des Bargelds	Politics	Jan. 13, 2018	187	9	24	393	1,160	20.78	7.79	2.10	6.20	442,794.23	898,932.76	4.08	1.76	342,680.51	745,545.09	3.94	1.74
	Gesundheit: Zahn-pasta könnte gegen Malaria helfen	Science	Jan. 19, 2018	179	10	14	366	1,080	17.90	12.79	2.04	6.03	451,792.01	862,434.12	4.01	1.93	371,003.02	771,694.36	3.82	1.91
	Belgien ist in der Silvesternacht geschrumpft: Grenze zu Nieder-landen wurde zu Neujahr korrigiert	Society	Jan. 2, 2018	189	14	19	389	1,184	13.50	9.95	2.06	6.26	378,199.22	765,992.15	4.09	1.72	328,029.85	671,520.36	3.90	1.79
	Deutschland enttäuscht bei Handball-EM: Nationalteam nur mit Remis gegen Mazedonien	Sports	Jan. 18, 2018	158	7	20	353	1,007	22.57	7.90	2.23	6.37	448,362.60	874,239.04	3.88	2.00	399,975.03	805,889.56	3.76	1.97
	Overall List B			181.29	10.86	19.00	392.57	1,125.29	17.35	9.90	2.17	6.21	428,902.94	26,891.67	3.94	0.14	368,955.57	26,318.52	3.80	0.14

CHAPTER III: ASSESSMENT OF TEXT COMPREHENSION

Tschense, M., & Wallot, S. (2022). Modeling items for text comprehension assessment using confirmatory factor analysis. *Frontiers in Psychology*, *13*:966347. https://doi.org/10.3389/fpsyg.2022.966347

Authorship status:	First author
Publication status:	Published on Oct 20, 2022
Scientific journal:	Frontiers in Psychology

The article is presented in its published version.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 October 2022
DOI 10.3389/fpsyg.2022.966347

Check for updates

OPEN ACCESS

EDITED BY
Philip Davis,
University of Liverpool,
United Kingdom

REVIEWED BY
Assis Kamu,
Universiti Malaysia Sabah,
Malaysia
Alexandr Nikolayevitch Kornev,
Saint Petersburg State Pediatric Medical University, Russia

*CORRESPONDENCE
Monika Tschense
monika.tschense@leuphana.de

SPECIALTY SECTION
This article was submitted to Psychology for Clinical Settings, a section of the journal Frontiers in Psychology

RECEIVED 10 June 2022
ACCEPTED 22 September 2022
PUBLISHED 20 October 2022

CITATION
Tschense M and Wallot S (2022) Modeling items for text comprehension assessment using confirmatory factor analysis. *Front. Psychol.* 13:966347. doi: 10.3389/fpsyg.2022.966347

COPYRIGHT
© 2022 Tschense and Wallot. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Modeling items for text comprehension assessment using confirmatory factor analysis

Monika Tschense[1,2,3]* and Sebastian Wallot[1,2]

[1]Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany, [2]Research Group for Research Methods and Evaluation, Institute of Psychology, Leuphana University Lüneburg, Lünebrug, Germany, [3]Research Group for Neurocognition of Music and Language, Planck Institute for Empirical Aesthetics, Frankfurt, Germany

Reading is a complex cognitive task with the ultimate goal of comprehending the written input. For longer, connected text, readers generate a mental representation that serves as its basis. Due to limited cognitive resources, common models of discourse representation assume distinct processing levels, each relying on different processing mechanisms. However, only little research addresses distinct representational levels when text comprehension is assessed, analyzed or modelled. Moreover, current studies that tried to relate process measures of reading (e.g., reading times, eye movements) to comprehension did not consider comprehension as a multi-faceted, but rather a uni-dimensional construct, usually assessed with one-shot items. Thus, the first aim of this paper is to use confirmatory factor analysis (CFA) to test whether comprehension can be modelled as a uni- or multi-dimensional concept. The second aim is to investigate how well widely used one-shot items can be used to capture comprehension. 400 participants read one of three short stories of comparable length, linguistic characteristics, and complexity. Based on the evaluation of three independent raters per story, 16 wh-questions and 60 yes/no-statements were compiled in order to retrieve information at micro and inference level, and 16 main contents were extracted to capture information at the macro level in participants' summaries. Still, only a fraction of these items showed satisfactory psychometric properties and factor loadings – a blatant result considering the common practice for item selection. For CFA, two models were set up that address text comprehension as either a one-dimensional construct (a uni-factor model with a single comprehension factor), or a three-dimensional construct reflecting the three distinct representational levels (three correlated first-order factors). Across stories and item types, model fit was consistently better for the three-factor model providing evidence for a multi-dimensional construct of text comprehension. Our results provide concrete guidance for the preparation of comprehension measurements in studies investigating the reading process.

KEYWORDS

reading, text comprehension, reading comprehension, comprehension assessment, discourse representation, mental model

Introduction

As we read, some kind of mental representation of the semantic structure of the text has to be generated, and – as long as reading progresses and new material (i.e., words) is processed – this model has to be expanded and updated constantly (Verhoeven and Perfetti, 2008; O'Brien and Cook, 2015).

As proposed by Kintsch and Van Dijk (1978), there are two levels to describe the semantic representation of a text, a local micro level and a more global macro level. The basic assumption is that every sentence of the text usually conveys at least one meaning (proposition). The micro level then refers to the whole set of propositions of the text, displaying only linear or hierarchical relations. However, the initial set of propositions has to be reduced and further organized in order to establish connections to the topic of discourse, but also to cope with cognitive limitations such as working memory capacity (Palladino et al., 2001; Radvansky and Copeland, 2001; Butterfuss and Kendeou, 2018). This results in a "meaningful whole" (Kintsch and Van Dijk, 1978, p. 366), a cohesive macro level of informational structure.

A third representation level, the so-called situation model or mental model, furthermore incorporates a reader's world knowledge and provides a scope for their own deductive and interpretive processes (Graesser et al., 1997; Van Den Broek et al., 2005; Sparks and Rapp, 2010). Thus, inferences can emerge that might exceed the literal meaning conveyed by a text (Perrig and Kintsch, 1985; Graesser et al., 1994, 1997). Since this theory considers both, first the construction of an (elaborated) propositional representation, and further the integration of readers' knowledge to form a final mental representation of a text, it is known as the construction-integration model (Wharton and Kintsch, 1991; Kintsch, 2005, 2018). While many more theories and models of text comprehension have been proposed, there is also a broad consensus that the representational structure described above is at the core of the vast majority of these theories and models (for a comprehensive review see McNamara and Magliano, 2009).

Previous research has found evidence that comprehension processes at each of these different levels are necessary (e.g., Perrig and Kintsch, 1985; Fletcher and Chrysler, 1990; McKoon and Ratcliff, 1992; Graesser et al., 1994; McNamara et al., 1996; Perfetti and Stafura, 2014; Kintsch, 2018; Lindgren, 2019), but there has been little research assessing comprehension at these different levels simultaneously. Moreover, current studies that investigated text comprehension in relation to process measures of reading did not assess and/or analyse comprehension scores according to different processing stages. For example, when factors such as text difficulty or inconsistencies and their effects on process measures of reading were investigated, comprehension was usually assumed but not explicitly tested (e.g., Rayner et al., 2006; for a review: Ferreira and Yang, 2019). Other studies relating the reading process to comprehension tried to assess comprehension by means of multiple-choice questions, but most of the time further information about how these items were compiled and/or which processing level they relate to were missing (e.g., LeVasseur et al., 2006, 2008; Wallot et al., 2014, 2015; O'Brien and Wallot, 2016). But even when different items for different processing levels were used (e.g., Schröder, 2011; Mills et al., 2017; Southwell et al., 2020), this differentiation was ultimately lost for further analyses due to averaging to uni-dimensional comprehension scores.

It should be noted that in none of the studies above pre-tests for item comprehensibility, difficulty or consistency were mentioned. It thus has is to be assumed that one-shot items were used in order to asses reader's text comprehension, relying heavily on the experimenters' intuition. With regards to post-hoc quality checks, Schröder (2011) was the only one implementing a comprehension evaluation by three independent raters, and was able to show a moderate level of inter-rater agreement (Fleiss' $\kappa = 0.64$). Furthermore, only Mills et al. (2017) included a reliability analysis and assessed the internal consistency of their comprehension items. However, this was a post-hoc analysis, and the resulting values for Cronbach's α ranged from 0.43 (unacceptable) to 0.86 (good) between texts, indicating high variability in item quality.

Looking at the respective findings, it is striking that in some of the referenced studies process measures of reading, e.g., reading times or eye movements, did relate to text comprehension (LeVasseur et al., 2008; Schröder, 2011; Southwell et al., 2020), but that these effects were lacking in others (LeVasseur et al., 2006; Wallot et al., 2015). Moreover, even when process measures were linked to participants' comprehension scores, effect sizes varied considerably depending on reading tasks (Wallot et al., 2014), data sets (Mills et al., 2017), or age groups (O'Brien and Wallot, 2016). Among the studies investigating the reading process in terms of self-paced reading, word reading speed generally did not predict comprehension well, often producing null-findings, while auto-correlation properties of the fractal scaling type of reading times fared somewhat better (Wallot et al., 2014, 2015; O'Brien and Wallot, 2016). Among the eye movement studies, the models predicting comprehension successfully did not do so based on the same process features (Wallot et al., 2015; Southwell et al., 2020). This state of affairs might be a question of how the reading process was modeled (i.e., which features of the reading process are of importance, and in which combination). However, the problem might also be the result of how the studies referenced above handled the measurement of reading comprehension.

All the studies mentioned above that tried to relate the reading process to comprehension seemed to have worked with one-shot items assessing comprehension through items with little to no systematic pretesting, and without establishing psychometric properties of these items before application. Moreover, they seemed to implicitly assume that comprehension is a uni-dimensional concept, with comprehension being mainly high or low (or present or absent) by averaging all items, or even using Cronbach's α as an indicator of reliability. However, to the degree that different levels at which comprehension can take place are distinguishable, a

TABLE 1 Participant demographics.

Short story	*N*	Sex			Age (years)			Reading per week (hours)		Educational level			
		Female	Male	Other	Range	*M*	*SD*	*M*	*SD*	Higher edu. entrance	Vocational qualification	Higher education	Other
1	117	93	24	0	[19, 77]	47.24	16.98	19.16	12.89	22	11	83	1
2	126	98	27	1	[19, 77]	46.42	14.32	20.38	12.41	13	16	91	6
3	140	111	28	1	[19, 91]	47.46	17.41	20.82	16.85	32	13	92	3
Overall	383	302	79	2	[19, 91]	47.05	16.29	20.17	14.31	67	40	266	10

Reading per week refers to the self-reported number of hours that participants approximately read per week (including books, newspaper articles, blog posts, etc.).

uni-dimensional concept might be misleading. The criticism raised here also applies to our own past work, which has followed the same practice and made the same assumptions (Wallot et al., 2014, 2015; O'Brien and Wallot, 2016). Accordingly, we are curious to find out, how good this practice of generating one-shot items can be in terms of producing reliable measures of comprehension, and in how far the assumption of uni-dimensionality is warranted in order to potentially improve future work.

Hence, the aim of the current study is to investigate how good the measurement properties of sets of one-shot comprehension questions are. Moreover, we aim to test whether and how items for comprehension assessment that target different levels of discourse structure (micro vs. macro vs. inference level) jointly contribute to text comprehension. For this purpose, we intend to deduce whether text comprehension can be measured and modelled as a uni-dimensional or multi-dimensional construct by means of confirmatory factor analysis (CFA). Additionally, as exploratory questions, we will investigate the relation between participants' text comprehension, their liking and interest ratings, as well as text reading times.

Materials and methods

The methods described below were approved by the Ethics Council of the Max Planck Society. Before inspection of any data, the study was preregistered *via* Open Science Framework (OSF[1]).

Participants

In total, 400 participants were recruited by distributing leaflets in local pedestrian zones, cafés, libraries, book stores and cinemas, placing advertisements at the institute's homepage and Facebook, as well as contacting participants *via* email using an in-house database and open email lists. At the end of the survey, participants could decide to join a lottery to win a book voucher of 10 € with odds of one in five. All participants were native speakers of German and at least 18 years old.

Two participants were excluded due to missing data of comprehension items and summary. Another 15 participants' data was excluded based on text reading times of less than 5 min or more than 40 min. Thus, the final sample consisted of 383 participants (302 females, 79 males, 2 others) with an age range between 19 and 91 years ($M = 47.05$, $SD = 16.29$). A majority of 69.45% of the participants stated holding a higher education degree. With regard to reading habits, participants reported to spend an average of 20.17 h per week ($SD = 14.31$) reading, for instance, books, newspaper articles, and blog posts. Participants were randomly assigned to one of three short stories, see Table 1 for distribution of demographic variables per text.

Materials

Texts

To allow for some generalization of the results across different texts, three short stories with different topics, but comparable complexity of content and pace of narration were selected. Short story 1 ("Brief an Juliane" [Letter to Juliane] by Hosse, 2009) describes the circumstances and challenges of growing up after World War II in an autobiographical manner (first-person narration). In contrast, short story 2 ("Die verborgene Seite der Medaille" [The hidden side of the coin] by Scavezzon, 2010) is a more typical short story with a third-person selective narrator and a plot twist towards its open end. Here, fact and fiction blend into one elaborate metaphor about the life of the main character, a veteran pilot that was involved in the bombing of Hiroshima. Short story 3 ("Der Doppelgänger" [The doppelganger] by Strauß, 2017) is a third-person omniscient narrative featuring a woman with Capgras syndrome, a psychological disorder leading her to the delusion that her husband has been replaced by an identical-looking impostor.

If necessary, the stories were adapted to current German spelling rules. Where possible, direct speech was either omitted or paraphrased. The texts were then shortened to a length of roughly 3,000 words to achieve a reading time of approximately 10–15 min (Brysbaert, 2019). The short stories were matched for number of words per sentence and mean length of words based on both, number of graphemes and number of syllables per word. Moreover, average logarithmic word frequencies obtained from

1 https://osf.io/2u43j

TABLE 2 Key characteristics per text.

Short story	Words	Sentences	Words per sentence	Graphemes per word	Syllables per word	Type frequency		Type frequency DC		Annotated type frequency	
						Absolute	log10	Absolute	log10	Absolute	log10
1	3,123	260	12.01	5.31 (2.99)	1.75 (0.96)	406,824.70	4.40	503,086.36	4.50	343,320.31	4.20
						(785,206.60)	(1.25)	(914,730.01)	(1.54)	(704,039.84)	(1.57)
2	2,967	244	12.16	5.02 (2.72)	1.69 (1.02)	371,672.56	4.56	445,139.65	4.66	318,950.96	4.38
						(695,293.86)	(1.32)	(78,6186.05)	(1.33)	(635,276.25)	(1.37)
3	3,113	262	11.88	5.29 (2.92)	1.77 (0.98)	398,567.54	4.47	505,960.92	4.57	337,254.16	4.30
						(749,976.33)	(1.44)	(961,725.28)	(1.45)	(673,702.76)	(1.47)

Words and sentences refer to the number of words and number of sentences per story, all other values are averaged per story; standard deviations are given in brackets.

dlexDB (Heister et al., 2011) were similar for all texts. See Table 2 for more information regarding text characteristics.

Comprehension items

To assess text comprehension as thoroughly as possible, different types of comprehension tasks were used. For each text, 60 yes/no-statements were generated, 40 of these aimed at micro-level content, the remaining 20 at inference-level content. Items assessing micro level comprehension related to information encoded at the sentence-level. Items assessing inferences did not have an explicit reference in the text as they exceed its literal meaning and integrate the reader's world knowledge. Here is an example:

> **Original text:**
> "Lore und ich verdienten uns unser Taschengeld dann beim Großbauern beim Erbsenpflücken, was damals noch per Hand gemacht wurde. Um sechs Uhr in der Frühe traf man sich und wurde zum Feld gekarrt. Zuweilen brannte die Sonne erbarmungslos, aber wir hatten ein Ziel. Wenn man fleißig war, hatte man am frühen Nachmittag einen Zentner, also fünfzig Kilogramm. Das war mühsam, denn Erbsen sind leicht. Man bekam dafür drei D-Mark, ein kostbarer Schatz, den man hütete."
> *[Lore and I then earned our pocket money by picking peas at a large farm, which was still done by hand at that time. We met at six in the mornings and were taken to the field. Sometimes the sun burned mercilessly, but we had a goal. If you were diligent, you got fifty kilograms by early afternoon. It was exhausting, because peas are light. In return we received three German marks, a precious treasure that we guarded.]*
> **Item for micro information:**
> Die Protagonistin half beim Erbsenpflücken, um sich Taschengeld zu verdienen.
> *The main character helped picking peas to earn some pocket money.*
> **Item for inferred information:**
> Die Protagonistin musste schon früh lernen, hart für ihr Geld zu arbeiten.
> *The main character had to learn early on to work hard for her money.*

Yes/no-statements provide a widely used and, with regards to procedure and analysis, fast and easy tool to evaluate text comprehension. However, in the absence of prior knowledge about such items, there is a risk of comparably high probability of guessing and the possibility that a certain context or wording may simplify giving the right answer. Therefore, 16 wh-questions with open input fields were compiled for each text, 10 of which for testing comprehension at micro level, the remaining six at inference level.

For both tasks, a larger pool of items was initially prepared with items either referring to a specific part of the story or relating to the overall plot. For yes/no-statements this initial item compilation consisted of 120 items per text, for wh-question an initial pool of 40 items was initially generated Supplementary Material 1. Subsequently, these items were independently judged by three raters. Finally, the best-rated 60 yes/no-statements and 16 wh-questions that were evenly distributed throughout the whole text were selected for data acquisition.

In order to examine text comprehension at macro level, three raters summarized the main contents of each story. Ideas that appeared in all three summaries were maintained; ideas that were mentioned in only two of the summaries were first discussed and subsequently either discarded or maintained. This resulted in 16 main ideas per text which were later on used to evaluate participants' summaries – i.e., counting the presence or absence of these ideas in each summary.

Procedure

An online study was set up using the platform SoSci Survey.[2] The study could be accessed from mid-December 2019 until mid-March 2020. At the beginning of the study, participants were informed about the aims and specific contents of the study, as well as data protection rules. Subsequently, they were asked for some socio-demographic information. Participants were then randomly assigned to one of the three short stories. They were instructed to

2 https://www.soscisurvey.de

read the assigned text in a natural manner, if possible, in quiet surroundings and without interruptions. The text was presented as a whole and participants could freely scroll up and down to go back or forth. The text was formatted in HTML with Arial font in size 3. Paragraphs were visually indicated with larger white space between lines. During the experiment, there was no set time limit for reading. On average, participants needed 12.97 min ($SD = 4.69$) to read a text.

After reading the short story, participants were required to write a brief summary reflecting the main contents of the short story. Subsequently, participants first answered the wh-questions followed by the yes/no-statements. All wh-questions were presented in one list but in randomized order. The sequence in which yes/no-statements were displayed was also randomized, and items were distributed across three pages of the survey. Finally, participants were asked to fill out a short questionnaire assessing their reading experience in terms of interest, liking, suspense, urgency, vividness, cognitive challenge, readerly involvement, rhythm, and intensity. To this end, participants were asked to rate how strongly they agree with a presented statement on a seven-point scale ranging from 0 ("not at all") to 6 ("extremely"). For the purpose of this study, we were only interested in participants' global interest ("How interested are you in the text?") and liking ("How much do you like the text?" "How gladly would you like to read similar texts?," "How strongly would you recommend the text to a friend?").

Item selection

Participants' answers to the wh-qhestions were assessed as true (1) or false (0). Furthermore, the written summaries were evaluated regarding the presence (1) or absence (0) of the 16 main ideas, thus, each summary could have received a maximum of 16 points. For this purpose, two raters familiarized themselves again with the text (i.e., reading the short story and reviewing its main ideas), and subsequently discussed and rated eight randomly drawn summaries together. The raters assessed another two summaries individually and then discussed their evaluations until they agreed upon a final assessment. This training was implemented to ensure best possible inter-rater reliability and took about 1.5 h per short story. Afterwards, both raters individually assessed all summaries corresponding to the respective short story (approximately 5.5 h per rater and text). The order of the summaries was randomized. Indeed, good inter-rater agreement was achieved as indicated by Krippendorf's α of 0.926 for short story 1, 0.936 for short story 2, and 0.902 for short story 3. Finally, discrepant evaluations were discussed until the raters agreed upon a final rating (roughly 1 h per text).

To filter out items with bad psychometric properties before computing any model, an item analysis was performed. As a first step, individual distributions of the items were inspected. Items that showed an accuracy rate of less than 5% or more than 95% were excluded from further analysis. Subsequently, joint distributions were observed by computing the phi coefficient (r_ϕ) for each pair of items. Since the different types of comprehension items are assumed to evaluate a different level of text comprehension, items of the same type are supposed to correlate with each other while items of different types should be not at all or less strongly correlated. Hence, items were successively excluded until items within a type reached an average r_ϕ between 0.1 and 0.9, and items between types did not exceed an average r_ϕ of 0.25.

With the remaining items a CFA was carried out using the R package *lavaan* (Rosseel, 2012). If the analysis did not converge, additional items were discarded based on their loadings, starting with the item with the lowest loading. When the analysis converged, standardized estimates were assessed and items with values of less than 0.2 and greater than 0.9 were removed.

Following the steps of the item analysis described above, at least three items for each item type could be retained per short story. An overview of the items can be found in Supplementary Material 2.

Results

The average reading time over all texts was 12.97 min ($SD = 4.69$), 15.08 min ($SD = 4.87$) for short story 1, 11.33 min ($SD = 4.14$) for short story 2, and 12.68 min ($SD = 4.36$) for short story 3. Participants' liking and interest ratings were in the medium range with an average score of 3.48 ($SD = 1.62$) respectively 3.68 ($SD = 1.54$) across all texts. For short story 1, ratings yielded an average of 3.51 ($SD = 1.64$) for likability and 4.02 ($SD = 1.56$) for interest. Short story 2 scored a mean likability rating of 3.68 ($SD = 1.61$) and a mean interest rating of 3.60 ($SD = 1.65$). For short story 3, mean likability was 3.27 ($SD = 1.60$) and mean interest was 3.46 ($SD = 1.36$). Regarding the comprehension items, participants average accuracy rates were 85.25% for yes/no-statements ($SD = 16.93$; short story 1: $M = 88.69\%$, $SD = 13.89$; short story 2: $M = 82.88\%$, $SD = 17.20$; short story 3: $M = 84.18\%$, $SD = 19.02$), 59.03% for wh-questions ($SD = 22.60$; short story 1: $M = 61.43\%$, $SD = 19.37$; short story 2: $M = 54.71\%$, $SD = 23.02$; short story 3: $M = 60.95\%$, $SD = 25.79$), and 53.87% for the main contents of the summaries ($SD = 29.38$; short story 1: $M = 41.35\%$, $SD = 28.03$; short story 2: $M = 66.82\%$, $SD = 27.22$; short story 3: $M = 53.46\%$, $SD = 28.85$). Accuracy rates per item are provided in Supplementary Material 2.

Comparing text comprehension models (CFA)

For each of the short stories, two different models were set up that reflect text comprehension as (A) one-dimensional construct implemented as uni-factor model with a single comprehension factor, or as (B) multi-dimensional construct capturing all levels of text comprehension (micro level, macro level, inferences) designed as a model containing three correlated first-order factors. All

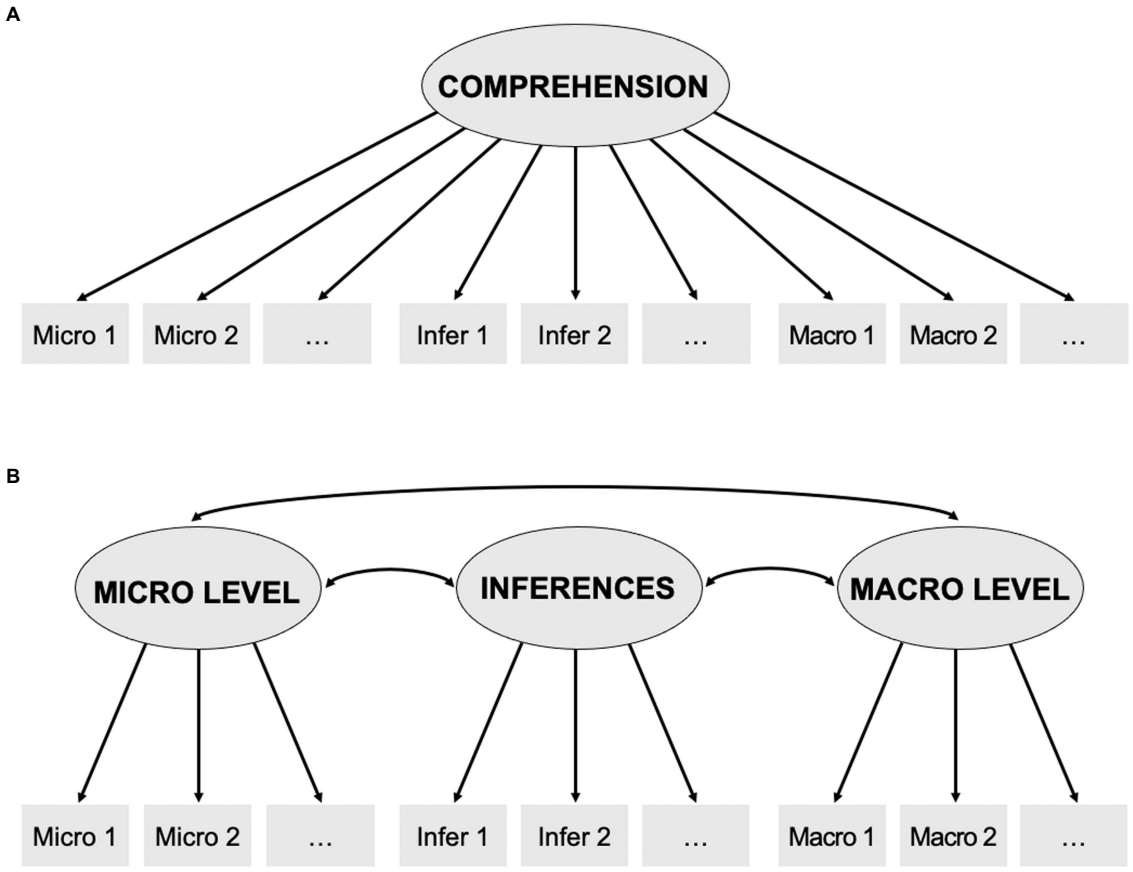


FIGURE 1
Models for confirmatory factor analysis (CFA). **A:** Uni-factor model of comprehension across three different items types (micro, macro, and inference level). **B:** Model of three correlated factors, assigning each item type its own latent variable.

models were conducted separately for wh-questions and yes/zo-statements. The specified models are shown in Figure 1. While we first planned to compute a third model based on the same multi-dimensional construct as in (B), extended by a second-order factor reflecting higher-level, general comprehension, this could not be realized due to converging errors.

Table 3 contains information about the goodness-of-fit indicators for each of the models described above. Both, unstandardized and standardized estimates are shown in Supplementary Material 3. When looking at yes/no-statements, model fit across all short stories is better for the three-factor model as compared to the uni-factor model. Turning towards the wh-questions, the same pattern emerges: Across all short stories, better model fit is indicated for the three-factor model than for the uni-factor model. When comparing the two types of comprehension tasks, some fit indices show even better model fit for wh-questions compared to yes/no-statements. Again, this pattern can be seen across all three short stories. In sum, the assumption that comprehension is a one-dimensional concept did not receive support from our model analysis. Note, that none of the models did converge when set up with the whole set of items; neither did the higher-order factor model.

Relation between comprehension, reading times, global interest and liking

In order to shed light on the relation between participants' comprehension scores, their ratings for global interest and liking of the text, as well as their reading times, Pearson's product–moment-correlation was computed for each pair of variables across short stories. To this end, reading time was logarithmized to adjust for normality, comprehension scores for the different discourse levels (micro vs. macro vs. inference level) were divided by their respective number of items, and an overall comprehension sum score was derived in the same manner, before all variables were z-transformed per short story. Results are shown in Table 4 for wh-questions, and in Table 5 for yes/no-statements.

As is evident in the correlation matrix, the different levels of text processing only show weak correlations among each other.

TABLE 3 Model fit per text.

Short story	Comprehension task	Model	ChiSQ				CFI	TLI	RMSEA			SRMR
			Value	df	ChiSQ / df	*p*			Value	90% CI	*p*	
1	Yes / no statements	A: uni-factor model	150.35	119	1.26	0.027	0.90	0.89	0.05	[0.017, 0.070]	0.549	0.21
		B: three-factor model	109.11	116	0.94	0.662	1.00	1.03	0.00	[0.000, 0.040]	0.989	0.18
	Wh-questions	A: uni-factor model	79.55	77	1.03	0.399	0.99	0.99	0.02	[0.000, 0.056]	0.905	0.15
		B: three-factor model	53.39	74	0.72	0.966	1.00	1.11	0.00	[0.000,0.000]	0.999	0.13
2	Yes / no statements	A: uni-factor model	166.73	152	1.10	0.196	0.91	0.90	0.03	[0.000, 0.051]	0.936	0.18
		B: three-factor model	103.46	149	0.69	0.998	1.00	1.30	0.00	[0.000, 0.000]	1.000	0.14
	Wh-questions	A: uni-factor model	116.63	90	1.30	0.031	0.78	0.74	0.05	[0.016,0.072]	0.516	0.15
		B: three-factor model	76.17	87	0.88	0.790	1.00	1.11	0.00	[0.000, 0.034]	0.994	0.12
3	Yes / no statements	A: uni-factor model	223.04	170	1.31	0.004	0.78	0.75	0.05	[0.028,0.064]	0.587	0.17
		B: three-factor model	153.68	167	0.92	0.762	1.00	1.06	0.00	[0.000, 0.028]	1.000	0.14
	Wh-questions	A: uni-factor model	69.89	77	0.91	0.705	1.00	1.06	0.00	[0.000, 0.038]	0.990	0.13
		B: three-factor model	50.25	74	0.68	0.984	1.00	1.18	0.00	[0.000, 0.000]	1.000	0.11

CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, root mean square error of approximation; SRMR, standardized root mean squared residual.

TABLE 4 Correlation matrix for wh-questions (selected items).

		Micro	Macro	Inference	Interest	Liking	Log reading time
Story 1	Micro	–	0.13	0.26**	−0.04	0.04	0.12
	Macro	0.13	–	0.23*	−0.06	0.01	0.08
	Inference	0.26**	0.23*	–	0.27**	0.26**	0.15
	Interest	−0.04	−0.06	0.27**	–	0.74***	0.09
	Liking	0.04	0.01	0.26**	0.74***	–	0.11
	Log reading time	0.12	0.08	0.15	0.09	0.11	–
Story 2	Micro	–	0.06	0.28**	0.09	0.09	0.13
	Macro	0.06	–	0.10	0.03	0.02	−0.09
	Inference	0.28**	0.10	–	−0.03	0.01	0.14
	Interest	0.09	0.03	−0.03	–	0.71***	0.03
	Liking	0.09	0.02	0.01	0.71***	–	0.02
	Log reading time	0.13	−0.09	0.14	0.03	0.02	–
Story 3	Micro	–	0.11	0.18*	0.01	−0.02	0.11
	Macro	0.11	–	0.04	0.04	0.10	0.09
	Inference	0.18*	0.04	–	0.12	0.23**	0.14
	Interest	0.01	0.04	0.12	–	0.68***	0.01
	Liking	−0.02	0.10	0.23**	0.68***	–	0.03
	Log reading time	0.11	0.09	0.14	0.01	0.03	–
Overall	Micro	–	0.10	0.24***	0.02	0.03	0.12*
	Macro	0.10	–	0.12*	0.01	0.05	0.03
	Inference	0.24***	0.12*	–	0.12*	0.17**	0.14**
	Interest	0.02	0.01	0.12*	–	0.71***	0.04
	Liking	0.03	0.05	0.17**	0.71***	–	0.05
	Log reading time	0.12*	0.03	0.14**	0.04	0.05	–

Pearson's *r* correlation coefficients. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

This is true for both, wh-questions and yes/no-statements. As could be expected, participants' global interest and liking of a short story are strongly correlated. However, a better reading experience does not relate to better comprehension of a text in a meaningful way. Furthermore, there is no strong evidence for a correlation between text comprehension and participants' reading times.

The pre-selection of comprehension items as described above descriptively leads to somewhat better discriminatory power between the three levels of text processing: There is a

TABLE 5 Correlation matrix for yes/no statements (selected items).

		Micro	Macro	Inference	Interest	Liking	Log reading time
Story 1	Micro	–	0.08	0.24**	0.03	0.14	0.14
	Macro	0.08	–	0.02	−0.03	0.00	0.02
	Inference	0.24**	0.02	–	−0.03	−0.02	0.01
	Interest	0.03	−0.03	−0.03	–	0.74***	0.09
	Liking	0.14	0.00	−0.02	0.74***	–	0.11
	Log reading time	0.14	0.02	0.01	0.09	0.11	–
Story 2	Micro	–	0.05	−0.04	0.11	0.10	0.18*
	Macro	0.05	–	−0.08	0.04	0.04	−0.07
	Inference	−0.04	−0.08	–	0.07	0.06	0.07
	Interest	0.11	0.04	0.07	–	0.71***	0.03
	Liking	0.10	0.04	0.06	0.71***	–	0.02
	Log reading time	0.18*	−0.07	0.07	0.03	0.02	–
Story 3	Micro	–	−0.03	0.11	−0.05	−0.02	0.12
	Macro	−0.03	–	0.07	0.03	0.08	0.12
	Inference	0.11	0.07	–	−0.10	−0.02	0.17*
	Interest	−0.05	0.03	−0.10	–	0.68***	0.01
	Liking	−0.02	0.08	−0.02	0.68***	–	0.03
	Log reading time	0.12	0.12	0.17*	0.01	0.03	–
Overall	Micro	–	0.03	0.10*	0.03	0.07	0.15**
	Macro	0.03	–	0.01	0.02	0.04	0.03
	Inference	0.10*	0.01	–	−0.02	0.01	0.09
	Interest	0.03	0.02	−0.02	–	0.71***	0.04
	Liking	0.07	0.04	0.01	0.71***	–	0.05
	Log Reading Time	0.15**	0.03	0.09	0.04	0.05	–

Pearson's *r* correlation coefficients. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

slight decrease in correlation coefficients for the selected items as compared to the whole item set. However, the overall relations between the investigated variables do otherwise remain the same. Correlation results for the whole item set across texts are displayed in Supplementary Material 4.

Discussion

The current study had two aims: First, we wanted to simultaneously model the three processing levels of comprehension (micro, macro and inference level). Particularly, we were interested in comparing a uni-factor model (i.e., that comprehension behaves the same across all of these three levels) with a model that assigns each of these levels their own factor. Second, we wanted to test the quality of different comprehension items in terms of capturing text comprehension after reading. This second point relates to the common practices of comprehension assessment, especially as applied in studies investigating the relation between process measures of reading and text comprehension. Here, researchers often seem to work with one-shot items of unknown psychometric quality, and to implicitly assume that comprehension is effectively a one-dimensional construct.

Our results indicated that a three-factor model of text comprehension fits our data significantly better than a uni-factor model. This was true for all three short stories and regardless of item type. Consequently, we provided evidence that comprehension should indeed be considered a three-dimensional construct. At the same time, our results showed that all three processing levels were correlated. This suggests three related, yet distinct levels of comprehension influencing one another. Thus, our analysis yields complementary evidence to studies investigating specific aspects of these processing levels separately. Accordingly, our results are in line with the assumption of three representational levels of discourse comprehension (micro, macro and inference level; *cf.* Kintsch and Van Dijk, 1978), also when these three levels are investigated simultaneously. In line with the theory, the results suggested a model with correlated factors, indicating that these levels are separate, but interdependent (*cf.* Perrig and Kintsch, 1985; Fletcher and Chrysler, 1990; McNamara et al., 1996; Perfetti and Stafura, 2014; Kintsch, 2018).

However, we would like to point out three aspects of our analysis that were somewhat striking. First, the standardized root mean squared residual (SRMR) values were quite high (≥0.11) for all models that converged, even though other fit indices were in the commonly expected range. Such larger SRMR values were reported before in the case of relatively small sample sizes of 200

or less due to higher degrees of uncertainty or variability that come along with smaller samples (*cf.* Taasoobshirazi and Wang, 2016). Second, when the whole initial item set was used in the comprehension models, none of the models converged. Thus, a comparison between the whole item pool and selected items was not possible indicating that items of poor and/or heterogenous quality are difficult to lump together into a single comprehension score. Third, it should be noted again that a higher-order factor model of text comprehension did not converge, indicating model misspecification. Even though this means we have no model fit indices to compare, it suggests that this is not an appropriate way to model the comprehension data.

As laid out in the introduction, it is currently common practice to assess comprehension in terms of one-shot items which are largely based on the experimenter's intuition for item selection than on theory, pre-tests or post-hoc quality control. As the current study showed, it is of importance to control comprehension items better, even if it requires quite some extra effort. The immense drop-out rate suggests that neither working with independent raters nor basing items on a theory by itself is enough to guarantee high item quality. Pre-testing items and/or reducing items post-hoc in a step-wise manner should be considered when planning further studies that aim to investigate text comprehension processes. Without investing some time and effort on item selection, there is a high risk that comprehension is not assessed in a valid manner and thus cannot be used in order to predict other measures of the reading process.

As we have summarized above, when we compared different studies relating reading process measures to comprehension, very different models emerge, and similar predictors behave differently across these studies (LeVasseur et al., 2006, 2008; Schröder, 2011; Wallot et al., 2014, 2015; O'Brien and Wallot, 2016; Mills et al., 2017; Southwell et al., 2020). This might be due to differences inherent in the specific reading situations (Wallot, 2016), but it might also be a function of varying quality of the comprehension assessment. Please note, that the current study was not a laboratory study, and accordingly, we had little control or information about the time course of reading behavior or the specific reading situation. Even though stricter experimental control is desirable in future work along these lines, this does not invalidate the main conclusion that can be drawn from our results: In order to draw reliable inferences about reading process measures that are related to reading comprehension, reliability and validity of comprehension measures is a necessary prerequisite. If the quality of comprehension measurements is unknown, however, it becomes difficult to trace back why a particular model of reading process measures was successful or failed in predicting reading comprehension as outcome.

Data availability statement

The dataset for this study is available in the online repository Open Science Framework (OSF): https://osf.io/b2zem/.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Council of the Max Planck Society. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MT designed the experiment and collected and analyzed the data. MT and SW jointly developed the research idea, contributed to the conceptualization of the study, interpreted the results, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) by grants to SW (project numbers 397523278 and 442405852).

Acknowledgments

We would like to thank Maria Raab, Franziska Roth and Nadejda Rubinskii for their help with stimulus selection, data collection and preprocessing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.966347/full#supplementary-material

References

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *J. Mem. Lang.* 109:104047. doi: 10.1016/j.jml.2019.104047

Butterfuss, R., and Kendeou, P. (2018). The role of executive functions in reading comprehension. *Educ. Psychol. Rev.* 30, 801–826. doi: 10.1007/s10648-017-9422-6

Ferreira, F., and Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Process.* 56, 485–495. doi: 10.1080/0163853X.2019.1591885

Fletcher, C. R., and Chrysler, S. T. (1990). Surface forms, textbases, and situation models: recognition memory for three types of textual information. *Discourse Process.* 13, 175–190. doi: 10.1080/01638539009544752

Graesser, A. C., Millis, K. K., and Zwaan, R. A. (1997). Discourse comprehension. *Annu. Rev. Psychol.* 48, 163–189. doi: 10.1146/annurev.psych.48.1.163

Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychol. Rev.* 101, 371–395. doi: 10.1037/0033-295X.101.3.371

Heister, J., Würzner, K. M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychol. Rundsch.* 62, 10–20. doi: 10.1026/0033-3042/a000029

Hosse, I. (2009). ABrief an Juliane. *In Doris Bock, Kurzgeschichten und Gedichte. Literarisches Fragment Frankfurt am Main. Norderstedt: Books on Demand, eds Hosse, I., Reimer, A., Saddai, A., Schön, G.* 54–72.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: the CI perspective. *Discourse Process.* 39, 125–128. doi: 10.1207/s15326950dp3902&3_2

Kintsch, W. (2018). "Revisiting the construction—Integration model of text comprehension and its implications for instruction," in *Theoretical Models and Processes of Literacy* (London, United Kingdom: Routledge), 178–203.

Kintsch, W., and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol. Rev.* 85, 363–394. doi: 10.1037/0033-295X.85.5.363

LeVasseur, V. M., Macaruso, P., Palumbo, L. C., and Shankweiler, D. (2006). Syntactically cued text facilitates oral reading fluency in developing readers. *Appl. Psycholinguist.* 27, 423–445. doi: 10.1017/S0142716406060346

LeVasseur, V. M., Macaruso, P., and Shankweiler, D. (2008). Promotion gains in reading fluency: a comparison of three approaches. *Read. Writ.* 21, 205–230. doi: 10.1007/s11145-007-9070-1

Lindgren, J. (2019). Comprehension and production of narrative macrostructure in Swedish: a longitudinal study from age 4 to 7. *First Lang.* 39, 412–432. doi: 10.1177/0142723719844089

McKoon, G., and Ratcliff, R. (1992). Inference during reading. *Psychol. Rev.* 99, 440–466. doi: 10.1037/0033-295X.99.3.440

McNamara, D. S., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cogn. Instr.* 14, 1–43. doi: 10.1207/s1532690xci1401_1

McNamara, D. S., and Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychol. Learn. Motiv.* 51, 297–384. doi: 10.1016/S0079-7421(09)51009-2

Mills, C., Graesser, A., Risko, E. F., and D'Mello, S. K. (2017). Cognitive coupling during reading. *J. Exp. Psychol. Gen.* 146, 872–883. doi: 10.1037/xge0000309

O'Brien, E. J., and Cook, A. E. (2015). "Models of discourse comprehension," in *The Oxford Handbook on Reading*. eds. A. Pollatsek and R. Treiman (New York: Oxford University Press), 217–231.

O'Brien, B. A., and Wallot, S. (2016). Silent reading fluency and comprehension in bilingual children. *Front. Psychol.* 7:1265. doi: 10.3389/fpsyg.2016.01265

Palladino, P., Cornoldi, C., De Beni, R., and Pazzaglia, F. (2001). Working memory and updating processes in reading comprehension. *Mem. Cogn.* 29, 344–354. doi: 10.3758/BF03194929

Perfetti, C., and Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Sci. Stud. Read.* 18, 22–37. doi: 10.1080/10888438.2013.827687

Perrig, W., and Kintsch, W. (1985). Propositional and situational representations of text. *J. Mem. Lang.* 24, 503–518. doi: 10.1016/0749-596X(85)90042-7

Radvansky, G. A., and Copeland, D. E. (2001). Working memory and situation model updating. *Mem. Cogn.* 29, 1073–1080. doi: 10.3758/BF03206375

Rayner, K., Chace, K. H., Slattery, T. J., and Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Sci. Stud. Read.* 10, 241–255. doi: 10.1207/s1532799xssr1003_3

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Scavezzon, B. (2010). Die verborgene Seite der Medaille. *In, Ein Album voller Kurzgeschichten. Frankfurt am Main: August-von-Goethe Literaturverlag*. 49–58.

Schröder, S. (2011). What readers have and do: effects of students' verbal ability and reading time components on comprehension with and without text availability. *J. Educ. Psychol.* 103, 877–896. doi: 10.1037/a0023731

Southwell, R., Gregg, J., Bixler, R., and D'Mello, S. K. (2020). What eye movements reveal about later comprehension of long connected texts. *Cogn. Sci.* 44:e12905. doi: 10.1111/cogs.12905

Sparks, J. R., and Rapp, D. N. (2010). Discourse processing—examining our everyday language experiences. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 371–381. doi: 10.1002/wcs.11

Strauß, J. (2017). "Der Doppelgänger." in Der Doppelgänger. *Psychiatrische Kurzgeschichten. Berlin, Heidelberg: Springer*, 1–18.

Taasoobshirazi, G., and Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, and TLI: an examination of sample size, path size, and degrees of freedom. *J. Appl. Quant. Methods* 11, 31–39.

Van Den Broek, P., Rapp, D. N., and Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Process.* 39, 299–316. doi: 10.1080/0163853X.2005.9651685

Verhoeven, L., and Perfetti, C. (2008). Advances in text comprehension: model, process and development. *Appl. Cogn. Psychol.* 22, 293–301. doi: 10.1002/acp.1417

Wallot, S. (2016). Understanding reading as a form of language-use: a language game hypothesis. *New Ideas Psychol.* 42, 21–28. doi: 10.1016/j.newideapsych.2015.07.006

Wallot, S., O'Brien, B. A., Coey, C. A., and Kelty-Stephen, D. (2015). "Power-law fluctuations in eye movements predict text comprehension during connected text reading," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. eds. D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock and C. D. Jenningset al. (Austin, TX: Cognitive Science Society), 2583–2588.

Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, H., and Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 1745–1765. doi: 10.1037/xlm0000030

Wharton, C., and Kintsch, W. (1991). An overview of construction-integration model: a theory of comprehension as a foundation for a new cognitive architecture. *ACM SIGART Bull.* 2, 169–173. doi: 10.1145/122344.122379

CHAPTER IV: REGULARITY MEASURES TO PREDICT TEXT COMPREHENSION

Tschense, M., & Wallot, S. (2023). *Using recurrence quantification analysis to measure reading comprehension for long, connected texts* [Manuscript submitted for publication]. Institute for Sustainability Education and Psychology, Leuphana University of Lüneburg.

Authorship status:	First author
Publication status:	Submitted on Sept 9, 2023
Scientific journal:	Journal of Experimental Psychology: Learning, Memory, and Cognition

Except for changes in layout and format, the manuscript is presented in its submitted version.

Abstract

Recent research that tried to link measures of the reading process (e.g., reading times, eye movements) to measures of the reading outcome (e.g., comprehension, fluency) produced heterogeneous results. After reviewing these findings and discussing possible reasons for it, we introduce the concept of reading time regularity (RTR; Tschense & Wallot, 2022a; Wallot, 2014; 2016) as a potential means to better capture the relationship between process and product measures of reading. Particularly, we tested whether regularity in the dynamics of eye movement behavior can predict text comprehension. To this end, 37 participants read fictional texts at a comfortable pace, as fast as possible (reading for speed), and as thoroughly as possible (reading for accuracy) while their eye movements were recorded. After reading, participants' text comprehension was assessed by yes-/no-statements and wh-questions. Recurrence quantification analysis (RQA) and sample entropy analysis (SampEn) were applied to gaze step data, and the number of fixations and fixation durations were extracted. The results show that recurrence measures but not SampEn successfully predicted comprehension. However, the effects were modulated by both, the item type used to assess comprehension, as well as the reading tasks. Furthermore, the number of fixations was found to predict comprehension, as was previously demonstrated by Southwell and colleagues (2020). Our findings suggest that measures of RTR complement these more traditional fixation-based approaches, but further research is needed to better understand the role of regularity as an indicator for the quality of the reading process.

Keywords: text reading, text comprehension, eye movements, recurrence quantification analysis (RQA), sample entropy analysis (SampEn)

It is generally assumed that reading comprehension is reflected in process measures of reading – usually in a form that is detrimental to reading-speed related components. That is, reading a "difficult" word takes longer than reading an "easy" word (e.g., length: New et al., 2006; frequency: Schotter et al., 2013), and reading a syntactically complex sentence takes longer than reading a syntactically simpler sentence (e.g., Kapteijns & Hintz, 2021; Staub 2010). Depending on the process measure at hand, the loss of reading speed can be due to qualitatively different reading patterns, for instance, a local increase in word reading times during a difficult passage, or an increased probability to re-read difficult passages (e.g., Rayner et al., 2006; Sturt 2007).

However, in the majority of studies that make a conceptual link between reading process and difficulty of linguistic processing, effects on comprehension are often not explicitly measured. Studies that investigated the relation between reading process measures (such as word reading times in self-paced reading or fixation durations during natural reading) have produced a very heterogenous picture of findings. In the following, we will briefly summarize this body of research and will discuss potential sources for this observed heterogeneity – which might lie in the conception and operationalization of reading process measures, but also in the reliability of comprehension measures involved. Then, we will introduce the concept of reading time regularity (RTR; Tschense & Wallot, 2022a; Wallot, 2014; 2016), which we think has some conceptual advantages as a gross-measure of the reading process compared to reading-speed related components. Finally, we will present results from an eye-tracking study investigating text reading with regard to different reading instructions that either emphasize speed or accuracy during reading.

The Relation of Linguistic Features and the Reading Process

A general assumption that all theories and models of reading share, is that the reading process is driven by linguistic features of written language – at least to some extent. This is particularly clear for the front-end of the reading process, visual word recognition, where lexical features (such as word frequency) have a substantial impact on word reading times (Grainger & Jacobs, 1996; Ziegler, Perry, & Coltheart, 2000). This assumption is instantiated in more encompassing models of the reading process where lexical and syntactic features govern fixation length, or the initiation/inhibition of a saccade during reading (Engbert et al., 2005; Reichle et al., 2009). However, this assumption is also shared by higher-level theories of reading that concern discourse comprehension, where higher-level linguistic features of a text, such as propositional density, situation model dimensions, syntactic complexity, etc. drive sentence reading times (Kintsch & Keenan, 1973; Graesser et al., 2004; Zwaan et al., 1995).

The problem of consistent effects of specific text features on reading process measures is highlighted by the fact that aspects of the reading process itself vary across different languages (Frost, 2012). Although certain aspects have been demonstrated across many languages, the impact of these factors diverges at the same time. For example, word frequency effects were found to be strong in English, but almost absent in Serbo-Croatian (Holden & Van Orden, 2002). Similarly, reading in many languages has been shown to be quite robust against changes in letter order, which has been described as a core property of reading at the neurophysiological level. Contrary to these findings, changes in letter order pose a great challenge for reading in Hebrew (Velan & Frost, 2007). Furthermore, such effects have been observed between reading tasks within the same language. For instance, the effect sizes for word frequency and word length are substantially larger for reading tasks that present isolated words or sentences compared to longer, connected text reading (Wallot et al., 2014;

Xiong et al., 2023). Effects of lexical features seem to decrease systematically across connected text reading (Wallot et al., 2013). Moreover, such effects can even entirely depend on the order in which reading tasks are performed: Teng and colleagues (2016) demonstrated that word frequency effects in a lexical decision task vanished when the lexical decision task followed connected text reading, but frequency effects were evident when the lexical decision task was performed first.

The Relation of Reading Process and Text Comprehension

At times, higher-level comprehension processes reveal themselves in process measures of reading. This is the case when comprehension fails, and the involved processes send feedback to perceptual levels of the reading process, for example, to pause (in order to reduce the rate of incoming information until comprehension problems are resolved), or to seek out specific parts of a text that were the source of comprehension problems. Here, comprehension problems effectively act as a disruption of the lawful relationship between linguistic text features that primarily govern word identification or sentence comprehension and reading process measures, for example, when a reader pauses for too long or slows down too much on a word in order to resolve comprehension problems (Blanchard & Iran-Nejad, 1987; Booth et al., 2018), or when a reader employs regressive eye movements to fixate previously read passages (Meseguer et al., 2002; Rayner et al., 2006).

Compared to the overall field of reading research, there are not many studies that investigate how process measures of connected text reading relate to reading comprehension, measuring both the reading process and reading comprehension on the same text materials. In a self-paced reading study, Schroeder (2011) showed a positive relationship between comprehension and reading speed related measures. Southwell and colleagues

(2020) reanalyzed data of three eye tracking experiments and found better text comprehension to be accompanied by more, but shorter fixations. Recent studies by Mézière and colleagues (2023a; 2023b) demonstrated that the predictive power of eye movement measures varied for different reading tasks and comprehension assessments. Other studies, however, failed to successfully relate measures of the reading process to reading comprehension (LeVasseur et al., 2006; 2008; Wallot et al., 2014; 2015). The reason for the difficulties to uncover such a linkage might be attributed to the varying ambiguous relationships between reading speed components and reading comprehension on the one hand (e.g., Wallot et al., 2014), as well as text features and reading process features on the other hand (Teng et al., 2016; Wallot et al., 2013).

Regarding the relation between reading speed and comprehension, it has long been established that high speed in reading isolated words is indicative of orthographic decoding mastery (Perfetti, 1985). Moreover, a positive correlation between speed in orthographic decoding tasks and tests for text comprehension could be demonstrated, suggesting speed and comprehension as components of general reading skill (Perfetti & Hogaboam, 1975; Jenkins et al., 2003). But there are also trade-off relations, for instance, when readers decrease reading speed in order to increase comprehension at difficult text passages (Carver, 1992; Dyson & Haselgrove, 2000), which could be interpreted as a lack of general reading skill. Nonetheless, utilizing this strategy requires reading-related skills such as meta-linguistic awareness (Zipke, 2007) or an understanding of discourse structure (Graesser et al., 1997). Dyslexic readers, who do not know where to look during reading, seem to lack such skills (Rayner, 1985). Hence, slowing-down strategies might rather be interpreted as a marker of good reading skill. Contrary to this, Breznitz and colleagues (2013) showed that forcing dyslexic children to read faster actually increased their comprehension. Moreover, research

on speed reading has provided evidence for the opposite effect, demonstrating a decrease of comprehension with increasing reading speed (Dyson & Haselgrove, 2000; Rayner, 1998).

A critical point that might explain some of the heterogeneous findings above is how reading comprehension was assessed. In a previous paper (Tschense & Wallot, 2022b), we reviewed studies that investigated the relation between reading process measures and comprehension, and noticed the common use of one-shot items with unknown psychometric properties for comprehension assessment. Moreover, all studies treated comprehension as a uni-dimensional concept, reflected in a single score to be used as dependent variable. After following a multi-step procedure to construct and select comprehension items, we still found the majority of them to not reliably measure text comprehension. In a sequence of tests using confirmatory factor analysis, we further found evidence that comprehension items did not conform to a unidimensional concept, but rather reflect different facets of comprehension.

Reading Time Regularity (RTR)

In the following, we introduce the concept of reading time regularity (RTR) as a general means to measure the coupling between linguistic information and perceptual-cognitive processes during reading (Wallot, 2014; 2016). Here, we argue that RTR has the advantage of inferring such coupling based solely on process measures (e.g., response times, eye movements etc.) without the need to specify how such performance is linked to specific text features. There is first evidence that recurrence-based measures can be used to capture informative changes in eye movements during reading. Based on series of gaze steps, lower recurrence measures could be associated with more disturbed reading conditions as well as control conditions not related to text reading (Tschense & Wallot, 2022a).

Furthermore, RTR might make for a conceptually clearer operationalization of reading fluency compared to reading speed components. RTR originally emerged to bridge the gap between process measures of text reading on the one hand, and text comprehension on the other hand. Given the theoretical background, we were specifically interested in defining measures of reading process data that correlate strongly with reading comprehension. Even though various measures of the reading process – such as word reading speed, fixation duration, amount of regressive eye movements etc. – have been shown to vary with local or global text difficulty (e.g., Just & Carpenter, 1980; Rayner et al., 2006), they perform only poorly in terms of predicting individual levels of reading comprehension (LeVasseur et al., 2006; 2008; Wallot et al., 2014; 2015).

While reading fluency is conceived as relatively effortless reading with at least average-to-good comprehension (O'Brien et al., 2014), reading fluency is often operationalized as overall reading speed or speed of reading time components. Here, level of speed is used as a stand-in measure for the reading process. However, reading speed can equally be seen as an outcome of reading ability. So far, this circularity issue in the presumed relationship of reading speed and comprehension is an empirically hard to avoid confound. As summarized above, the relationship between reading speed and comprehension is complex: While speed is thought to correlate positively with comprehension as a general aspect of reading ability, increasing reading speed can lead to both, increased but also decreased comprehension trade-off-relationships. Therefore, adding the concept of RTR into an operational definition of reading fluency might be able to resolve this conceptual problem. When RTR is used as a measure for reading process fluency in the sense of an effortful, functional execution of the reading process, speed can be solely treated as an outcome variable. In a previous study,

measures of RTR have shown a predictive link to reading speed and comprehension, and captured their trade-off relation well (Wallot et al., 2014).

One source of this problem seems to lie in the relationship between specific linguistic text features (e.g., lexical word properties or syntactic constructions) and how they are operated on by the perceptual-cognitive processes during reading. Generally speaking, linguistic text features are thought to significantly co-control perceptual-cognitive processes during reading, for example determining how long a reader fixates a word before initializing a saccade to the next word (Engbert et al., 2005; Reichle et al., 2009). This also means that variations of such features lead to increased (or decreased) processing difficulty during reading, which in turn can have consequences for reading comprehension. The problem is, however, that the coupling between reader performance and specific linguistic text features is highly variable across individuals (Rayner et al., 2006; Traxler et al., 2012), tasks (Teng et al., 2016; Wallot et al., 2013), and languages (Holden & Van Orden, 2002; Frost, 2012).

Since the calculation of RTR does not depend on specific linguistic text features, it could in principle be used as a cross-linguistic measure for the prediction of reading comprehension, irrespective of the particular properties of different writing systems and their consequences for reading. Prior work using measures of regularity of the reading process has shown that the degree of regularity in reading time data is predictive of reading comprehension. Notably, RTR properties reliably predicted text comprehension better than reading speed (O'Brien & Wallot, 2016; O'Brien et al., 2014; Wallot et al., 2014). Preliminary results from an eye tracking study corroborated the power of RTR measures in predicting text comprehension over and above standard eye movement features, such as average fixation duration, number of fixations, and percentage of regressive eye movements (Wallot et al., 2015).

A basic proposal for the relationship between linguistic text features, reading process measures and reading comprehension could be formulated as follows:

(A1) There is a systematic relationship between linguistic text features and the reading process, where linguistic features explain variance of an observable of the reading process (e.g., word frequency predicts word recognition time).
(A2) There is a systematic relationship between observables of the reading process and text comprehension (e.g., fixation duration predicts comprehension).
(A3) There is a systematic relationship between (i) the strength of the correlation between reading process measures and linguistic text features and (ii) text comprehension (e.g., the correlation strength of fixation durations with word frequency predicts comprehension).

Operationally, RTR of a reading process measure can in principle be calculated by any statistic that captures order of a sequence/time-series (such as recurrence quantification analysis (Zbilut & Webber, 1992) or sample entropy (Richman & Moorman, 2000) – which we briefly describe in the methods section. The fact that RTR is solely based on the values of an observable of the reading process – and not particular text features – can address the challenges outlined above: Because RTR is independent of reading speed, its relationship to comprehension is not burdened by the same trade-off relations summarized above and RTR might be very well suited as a process measure of reading fluency.

Aims of the Current Investigation

In the current study, participants read fictional texts following three different reading instructions: (A) reading the text as fast as possible, while still retaining a minimum level of comprehension (reading for speed); (B) reading the text at a comfortable pace; and (C) reading the text as accurately as possible in order to reach a maximum level of comprehension (reading for comprehension). We aim to test whether regularity measures of eye movements

are predictive of reading comprehension. Following earlier findings (O'Brien & Wallot, 2016; O'Brien et al., 2014; Tschense & Wallot, 2022a; Wallot et al., 2014; 2015), higher recurrence properties and entropy rates should predict better reading comprehension (***Hypothesis 1***).

Moreover, we were interested to investigate whether the effects of regularity measures on comprehension are influenced by the different reading conditions, or invariant across them. Ideally, effects of regularity measures are invariant across reading conditions (i.e., reading for speed, reading for comprehension, or reading at a comfortable pace; ***Hypothesis 2***). This is the strong invariance hypothesis, showing only a main effect of regularity measures on comprehension, but no main effect of condition and no interaction. Alternative one represents a weak invariance of regularity measures, meaning that there is both, a main effect of regularity measures on comprehension, and a main effect of reading condition. This implies, that there is additional variance in comprehension measures that cannot be proximally explained by regularity alone, and conditions might differ in their levels of recurrence and entropy measures or comprehension scores. Alternative two depicts uncertain invariance of regularity measures, where the effect of regularity measures on comprehension depends on reading condition.

Additionally, we attempt to conceptually replicate a model by Southwell and colleagues (2020), where more, but shorter fixations were associated with better reading comprehension (***Hypothesis 3***). Finally, we compare the more traditional fixation-based approach and RTR (***Exploration***).

Method

The study as described below was approved by the Ethics Council of the Max Planck Society and followed the ethical principles of the Declaration of Helsinki. Before any data was inspected and analyzed, the study was preregistered via the Open Science Framework (OSF; https://osf.io/96hb8). The data and analysis scripts associated with this manuscript can be accessed here: https://osf.io/2h9pr.

Participants

Forty-five native speakers of German with normal or corrected-to-normal vision took part in the study. All participants were required to be between 18 and 60 years old, to not have a reading disorder or other psychological disabilities. They received a compensation of 7€ per half hour up to a maximum of 21€. Due to problems during the calibration procedure, eight participants dropped out of the study, and their data was excluded from analysis. Thus, the final sample consisted of 37 participants (26 female, 11 male) between 21 and 59 years of age (*M* = 31.78 years, *SD* = 11.24). Participants indicated to read on average 21.11 hours per week (*SD* = 13.06), however, self-reports varied from as little as 3.50 to a maximum of 50.00 hours per week. Written informed consent was obtained prior to participation.

Stimuli

All stimuli used here were tested in Tschense and Wallot (2022b); we refer the reader there for detailed information on the selection procedure. Key details about the stimuli are outlined below. Three German short stories with different topics but of comparable plot complexity were selected. The texts were matched for length (i.e., number of words and sentences), average word length (i.e., number of graphemes and syllables per word), and average word

frequency (**Table 1**). As tested previously, the average reading time across stories was 12.97 minutes (*SD* = 4.69), and participants' likability and interest ratings were in the medium range on a seven-point scale (likability: *M* = 3.48, *SD* = 1.62; interest: *M* = 3.68, *SD* = 1.54).

Table 1

Key characteristics per text

Short Story	Words	Sentences	Graphemes per Word	Syllables per Word	Type Frequency		Annotated Type Frequency	
					absolute	log10	absolute	log10
1	3123	260	5.31 (2.99)	1.75 (0.96)	406824.70 (785206.60)	4.40 (1.25)	343320.31 (704039.84)	4.20 (1.57)
2	2967	244	5.02 (2.72)	1.69 (1.02)	371672.56 (695293.86)	4.56 (1.32)	318950.96 (635276.25)	4.38 (1.37)
3	3113	262	5.29 (2.92)	1.77 (0.98)	398567.54 (749976.33)	4.47 (1.44)	337254.16 (673702.76)	4.30 (1.47)

Note. Words and sentences refer to the number of words and number of sentences per story, all other values are averaged per story; standard deviations are given in brackets. Frequency values were obtained from dlexDB (Heister et al., 2011).

For each of the texts, a battery of comprehension items was compiled consisting of yes/no-statements and wh-questions (**Supplement 1**). Items of both types refer to either micro information literally mentioned in the text, or inferences arising during reading. For the yes/no-statements, participants read a statement related to the text and subsequently had to decide whether the presented information was true or false. The wh-questions could be openly answered with words or word groups. All items were subsequently translated into correct or wrong answers coded as "1" or "0", allowing us to compute overall comprehension scores.

Procedure

Participants were comfortably seated in a soundproof both with dimmed light. Supported by a head and chin rest, they looked at an LCD-monitor (size: 24 in, refresh rate: 144 Hz, resolution: 1920 x 1080 px). The distance between eyes and monitor was 70 cm. An

EyeLink 1000 (SR Research) was used for binocular recording of eye movements at a sampling rate of 500 Hz. The study was implemented in OpenSesame (version 3.3.5; Mathôt et al., 2012) using the EyeLink display software (SR Research) and the PyGaze toolbox (Dalmaijer et al., 2014). On the screen, instructions and stimuli were presented in white on black background. Text was displayed left-aligned in monospaced font with a size of 40 px.

The study was conducted in one session of approximately 90 minutes, varying according to participants' individual reading speed. Every participant read all three short stories and answered comprehension questions prompted immediately after each text. Between texts, reading instructions were manipulated: Participants were asked to read one text at a comfortable pace, one as quickly as possible, and one as accurately as possible. Both, the order of texts and the order of reading instructions were randomized. A 12-point calibration in random sequence followed by a validation procedure preceded each reading block. Between blocks, participants were allowed to take a short break.

At the beginning of each text, a fixation cross was displayed at the top-left, indicating the location of the first word. When participants fixated the cross, the first page of the short story appeared on the screen. Participants proceeded to the next page or, at the end of the text, to the comprehension battery in a self-paced manner by pressing the space bar. Following each short story, participants first answered the corresponding wh-questions. They were instructed to type in brief answers via the keyboard. Subsequently, participants judged the respective yes-/no-statements as true or false via mouse click. All comprehension items were presented in randomized order, one item at a time.

Data Analysis

Participants' reading times were extracted per trial and divided by the word count of the respective text, resulting in average reading times per word. Based on participants' average response accuracy, comprehension scores were calculated separately per item type and trial. Additionally, comprehension scores were normalized by word reading times in a trial-wise manner (comprehension-reading time-ratios; cf. Wallot et al., 2014).

The recorded eye movement data were used to extract gaze steps and fixation measures as dependent variables. Gaze steps were computed by differencing the raw two-dimensional position data of consecutive samples per trial (Stephen & Mirman, 2010). Extreme values deviating more than 25 *SD* from the mean were discarded. Based on the Microsaccade Toolbox for R (Engbert et al., 2015), number of fixations and fixation durations were calculated per trial, setting the minimal number of samples per saccade to 6 and specifying a velocity factor of 5. Subsequently, fixation durations were averaged per trial. The number of fixations per trial was divided by the total number of words in order to account for differences in text lengths.

A priori, the following drop-out criteria were specified: Participants with reading comprehension substantially below chance level were to be excluded, and data sets in which more than 10% of data points per trial were defective (e.g., due to artefacts, blinks) were to be discarded. Please note that none of the participants had to be excluded based on either chance-level comprehension or erroneous eye movement data.

Recurrence Quantification Analysis (RQA)

RQA can be used to quantify various dynamic properties of a time series related to the degree of randomness and structure of its temporal evolution. It can be visualized by means of

recurrence plots (RP) based on which several complexity measures can be derived quantifying the density of recurrence points and their line structures (Marwan et al., 2007, Wallot, 2017).

Several RQA measures can be extracted from an RP, but we will focus on the following ones: The recurrence rate (RR) refers to the density of recurrence points, providing information about the repetitiveness of individual values or coordinates within a timeseries.

To the degree that the dynamics of a time series are coordinated in terms of temporally extending states, spanning multiple values, individual recurrence points more often occur adjacent to each other, producing clustered recurrence. For highly stochastic data that we are investigating, further recurrence measures can be used to quantify the degree of regularity in a time series, namely percent laminarity (LAM) and trapping time (TT). LAM equals the percentage of individual recurrences that are part of a recurrence cluster. TT captures the average size of those clusters, which is equal to the duration that the dynamics of the time series are "trapped" within a single cluster. Further measures indicate deterministic trajectories in a time series, which refers to connected recurrence points within a RP. The determinism rate (DET) is the ratio of recurrent points occurring in connected trajectories to all recurrent points of a time series. Consecutive recurrences create lines structures, which are typically specified in terms of the average length of diagonal lines (ADL) and the maximal length of diagonal lines (MDL; Wallot, 2017).

Before running RQA, a delay parameter τ, and an embedding parameter D have to be estimated. Here, a delay parameter $\tau = 10$ and an embedding parameter $D = 10$ were estimated based on the average mutual information and false nearest neighbor functions. Z-scored series of gaze steps were then subjected to RQA (Zbilut & Webber, 1992) using the crqa package for R (v2.0.3; Coco et al., 2021). Due to computational limits, RQA for gaze step data was performed in a windowed manner (window size: 10,000; step: 5,000), and then

averaged per trial. The threshold parameter $T = 0.80$ was selected in an iterative procedure, aiming to obtain an average RR between 5% and 10% across the whole sample of trials and participants (Wallot, 2017). The chosen set of parameters resulted in a mean RR of 6.56% ($SD = 3.52$).

Sample Entropy Analysis (SampEn)

SampEn quantifies the degree of predictability of a timeseries (Richman & Moorman, 2000). It takes into account the number of matching sequences identified within a tolerance band defined by a radius *r* (excluding self-matches). Specifically, SampEn is the average probability that a sequence with length of $m + 1$ data points finds a matching sequence within radius *r*, given that a match for *m* data points has already been found (for a tutorial, see Kuznetsov et al., 2013). Highly periodic, deterministic timeseries are easily predictable (i.e., if sequences of *m* points repeat, then sequences of $m + 1$ points are also likely to repeat), yielding SampEn = 0. In contrast, a timeseries that is very noisy yields SampEn > 0.

SampEn was computed based on gaze step and fixation duration data using a custom script in MATLAB (vR2020b; MathWorks Inc., 2020). In general, SampEn is calculating the number of matching sequences of some length *m* and $m + 1$ within a tolerance band defined by a radius *r* (Richman & Moorman, 2000). Following an approach proposed by Ramdani and colleagues (2009), we chose a template length of $m = 1$ and specified a tolerance region of $r = 3$.

Inferential Statistics

As for our first hypothesis, recurrence measures (RR, DET, LAM, ADL, MDL, TT) and SampEn were used to predict participants' reading comprehension (comprehension scores and comprehension-reading time-ratios). Additionally, the effect of condition was

investigated in order to test hypothesis two. To replicate results by Southwell and colleagues (2020), average fixation durations and number of fixations were predictors for our third hypothesis. As an exploratory analysis, we also tested how both, regularity- and fixation-based measures together performed as predictors for comprehension. We set up mixed-effects models in RStudio (v4.2.0; R Core Team, 2022) using the lme4 package (v1.1-30; Bates et al., 2015), and tested for statistical significance using the lmerTest package (v3.1-3; Kuznetsova et al., 2017). All models were specified according to the following general form:

$$y_{mi} = y_{00} + y_{01}MEASURE_{mi} + \upsilon_{0i} + \varepsilon_{mi}, \quad \varepsilon \sim N(0, \sigma^2)$$

Here, y_{00} is the fixed intercept, $y_{01}MEASURE_{mi}$ is the fixed effect of the measure(s) of interest, υ_{0i} is the random intercept for participants, and ε_{mi} is the error term. While the fixed effect $y_{01}MEASURE_{mi}$ remains unchanged for the assumption of strong invariance of hypothesis two, it is complemented by a fixed effect for condition ($y_{01}MEASURE_{mi} + y_{01}CONDITION_{mi}$) to test for weak invariance, and substituted by an interaction term ($y_{01}MEASURE_{mi} * y_{01}CONDITION_{mi}$) to reflect uncertain invariance.

Results

On average, participants spent 12.63 minutes (*SD* = 3.79) on reading each text, and afterwards answered 72.80% (*SD* = 14.51) of the yes-/no-statements as well as 54.31% (*SD* = 23.67) of the wh-questions correctly. Trial duration was shortest for the fast reading condition (*M* = 9.22 min, *SD* = 1.89) which coincided with the lowest comprehension scores (yes-/no-statements: *M* = 66.22%, *SD* = 14.56; wh-questions: *M* = 54.31%, *SD* = 23.67). The normal reading condition resulted in intermediate trial duration (*M* = 13.76 min, *SD* = 3.44) along with intermediate comprehension scores (yes-/no-statements: *M* = 75.60%, *SD* = 13.84; wh-questions: *M* = 60.04%, *SD* = 21.21). As for the slow reading condition, trial duration was longest (*M* = 14.90 min, *SD* = 3.13) and coincided with the highest comprehension scores (yes-/no-statements: *M* = 76.58%, *SD* = 13.14; wh-questions: *M* = 62.01%, *SD* = 19.87). More information is provided in **Table 2**.

Table 2

Reading times and text comprehension scores

Condition	Trial Duration [min]		Word Reading Time [ms]		Comprehension Score				Comprehension-Reading Time-Ratio			
					yes/no		wh		yes/no		wh	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
overall	12.63	3.79	246.82	72.90	72.80	14.51	54.31	23.67	0.315	0.094	0.232	0.114
normal	13.76	3.44	269.01	63.72	75.60	13.84	60.04	21.21	0.293	0.071	0.240	0.105
fast	9.22	1.89	179.67	35.98	66.22	14.56	40.89	24.23	0.379	0.098	0.232	0.139
slow	14.90	3.13	291.76	60.56	76.58	13.14	62.01	19.87	0.274	0.078	0.224	0.094

Note. Word reading time was calculated as trial duration divided by the number of words; *yes/no* refers to the yes-/no-statements, and *wh* to the wh-questions used to assess text comprehension after reading.

Descriptive statistics for all measures that were extracted based on eye movement recordings are shown in **Table 3**. As can be seen in **Table 4**, recurrence measures and SampEn are highly intercorrelated, reflecting that they all capture the concept of regularity exhibited in gaze step data. As indicated by Bartlett's test of sphericity ($\chi^2(15) = 1470.20$, $^{***}p < .001$) and the Kaiser-Mayer-Olkin index (overall KMO = 0.70, KMO for all variables > 0.64), data were suitable for reduction. Thus, principal component analysis (PCA) was carried out using the psych package for R (v2.3.3; Revelle, 2023). Based on a parallel analysis, one principal component could be determined that reflects regularity (see **Supplement 2**).

Hypothesis 1: Recurrence Measures and SampEn as Predictors for Comprehension

To test for associations between reading comprehension and recurrence and entropy properties of the time series, linear mixed-effects models were estimated separately for each of the regularity measures, the reduced regularity component, and SampEn. Estimations were conducted separately for comprehension measure (comprehension scores vs. comprehension-reading time-ratios) and item type (yes-/no-statements vs. wh-questions). Results for all computed models are summarized in **Table 5** and **Table 6** below.

Table 3

Descriptive statistics for fixation measures, recurrence measures and SampEn

Condition	Number of Fixations		Number of Fixations / Words		Fixation Duration		RR		DET		MDL		ADL		LAM		TT		SampEn	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
overall	2780.54	807.02	0.9058	0.2592	219.04	17.40	6.56	3.52	84.81	9.66	401.96	362.55	4.01	0.96	91.7280	6.2302	5.5897	1.6220	0.0587	0.0208
normal	3021.27	741.83	0.9841	0.2285	221.13	17.76	7.16	3.49	85.24	9.61	411.63	360.78	4.06	0.94	91.8571	6.2539	5.7086	1.6061	0.0570	0.0206
fast	2081.49	421.54	0.6762	0.1343	212.71	16.51	4.67	2.17	83.43	9.97	339.73	350.33	3.80	0.92	91.1577	6.5166	5.1673	1.5219	0.0639	0.0217
slow	3238.86	695.94	1.0570	0.2258	223.26	16.53	7.84	3.90	85.77	9.50	454.51	376.61	4.16	0.99	92.1691	6.0393	5.8933	1.6884	0.0552	0.0197

Note. RR: Recurrence rate; DET: Determinism rate; MDL: Maximum diagonal line length; ADL: Average diagonal line length; LAM: laminarity; TT: Trapping time; SampEn: Sample entropy.

Table 4

Correlation matrix for recurrence measures

	RR	DET	MDL	ADL	LAM	TT	SampEn
RR	–	0.62	0.58	0.72	0.56	0.77	-0.55
DET	0.62	–	0.62	0.77	0.99	0.78	-0.55
MDL	0.58	0.62	–	0.88	0.54	0.87	-0.82
ADL	0.72	0.77	0.88	–	0.69	1.00	-0.75
LAM	0.56	0.99	0.54	0.69	–	0.70	-0.46
TT	0.77	0.78	0.87	1.00	0.70	–	-0.75
SampEn	-0.55	-0.55	-0.82	-0.75	-0.46	-0.75	–

Note. Pearson's r correlation coefficients; all coefficients are significant at the $p < .001$ level.

Table 5

Results of mixed-effects models: Comprehension scores predicted by regularity measures

Measure	Yes-/no-statements			Wh-questions		
	χ^2	*df*	*p*	χ^2	*df*	*p*
RR	1.89	1	.169	2.64	1	.104
DET	0.06	1	.800	2.02	1	.156
MDL	0.05	1	.820	0.14	1	.706
ADL	0.01	1	.915	0.04	1	.834
LAM	0.13	1	.716	2.30	1	.130
TT	0.01	1	.909	0.00	1	.979
Regularity	0.08	1	.777	0.01	1	.934
SampEn	0.79	1	.374	0.51	1	.473

Note. Comprehension scores were used as dependent variable; models were compared against a null model containing only random effects. Regularity: determined first principal component.

Table 6

Results of mixed-effects models: Comprehension-reading time-ratios predicted by regularity measures

Measure	Yes-/no-statements				Wh-questions			
	χ^2	*df*	*p*		χ^2	*df*	*p*	
RR	27.65	1	<.001	***	7.75	1	.005	**
DET	2.21	1	.137		6.03	1	.014	*
MDL	0.01	1	.912		0.00	1	.949	
ADL	2.16	1	.142		4.15	1	.042	*
LAM	1.64	1	.200		5.14	1	.023	*
TT	3.36	1	.067	.	4.63	1	.031	*
Regularity	4.29	1	.038	*	5.02	1	.025	*
SampEn	0.33	1	.566		0.82	1	.366	

Note. Comprehension-reading time-ratios were used as dependent variable; models were compared against a null model containing only random effects. Regularity: determined first principal component. *p < .05, **p < .01, ***p < .001.

Regardless of the item type, none of the regularity measures significantly predicted participants' comprehension scores. Turning towards comprehension-reading time-ratios, RR consistently predicted comprehension scores across item types (yes-/no-statements: $\chi^2(1) = 27.65$, ***$p < .001$; wh-questions: $\chi^2(1) = 7.75$, **$p = .005$). The same effect was evident for the reduced regularity component (yes-/no-statements: $\chi^2(1) = 4.29$, *$p = .038$; wh-questions: $\chi^2(1) = 5.02$, *$p = .025$). TT yielded a significant effect for wh-questions ($\chi^2(1) = 4.63$, *$p = .031$), however, only a trend was found for yes-/no-statements

($\chi^2(1) = 3.36$, $.p = .070$). Further significant effects for wh-questions occurred for DET ($\chi^2(1) = 6.03$, $*p = .014$), ADL ($\chi^2(1) = 4.15$, $*p = .042$) and LAM ($\chi^2(1) = 5.14$, $*p = .023$).

While regularity measures failed to predict comprehension scores, we did find effects for comprehension-reading time-ratios. However, not all regularity measures successfully predicted participants' comprehension ratios, and merely RR and the regularity component (and TT) did so across both item types. Furthermore, the direction of effects seemed to contradict our hypothesis: higher comprehension ratios corresponded to less regularity.

Hypothesis 2: Strong vs. Weak vs. Uncertain Invariance of Effects

To further investigate the relationship between comprehension and regularity, three mixed-effects models were set up reflecting a strong, weak and uncertain invariance of effects according to our second hypothesis. All models were estimated separately per regularity measure and comprehension item type. The models were compared to a null model containing only the random-effects structure. Results are provided in **Table 7**. Based on the findings above, all following results were based on comprehension-reading time-ratios; results for comprehension scores are provided in **Supplement 3**.

Again, RR showed a consistent pattern of results across both comprehension item types in line with the assumption of strong invariance of effects. However, all other regularity measures regarding yes-/no-statements provided support for a weak invariance of effects driven by the additional fixed factor of reading condition. Turning towards wh-questions, we only obtained evidence supporting a strong invariance of effects (RR, DET, ADL, LAM, and TT) or none of the invariance assumptions (MDL). Thus, the relationship between comprehension and regularity was not affected by reading condition in this case. SampEn patterns reflected weak invariance for yes-/no-statements, but did not yield any effects for wh-questions. In summary, our results suggested mixed evidence mostly in support of a strong invariance assumption regarding wh-questions, but rather in line with a weak invariance assumption for yes-/no-statements. This pattern of results was corroborated for the regularity component.

Table 7

Model comparison: Invariance of effects

Measure	Model		Yes-/no-statements							Wh-questions						
		nPar	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*
RR	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-230.55	-219.71	119.28	-238.55	27.65	1	<.001***	-197.76	-186.93	102.88	-205.76	7.75	1	.005**
	weak	6	-245.11	-228.85	128.56	-257.11	18.56	2	<.001***	-197.87	-181.62	104.94	-209.87	4.11	2	.128
	uncertain	8	-241.66	-219.98	128.83	-257.66	0.55	2	.760	-194.50	-172.82	105.25	-210.50	0.62	2	.733
DET	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-205.12	-194.28	106.56	-213.12	2.21	1	.137	-196.04	-185.21	102.02	-204.04	6.03	1	.014*
	weak	6	-234.93	-218.67	123.46	-246.93	33.81	2	<.001***	-192.92	-176.66	102.46	-204.92	0.87	2	.646
	uncertain	8	-233.28	-211.61	124.64	-249.28	2.35	2	.308	-188.92	-167.24	102.46	-204.92	0.00	2	.999
MDL	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-202.91	-192.08	105.46	-210.91	0.01	1	.912	-190.02	-179.18	99.01	-198.02	0.00	1	.949
	weak	6	-234.74	-218.48	123.37	-246.74	35.83	2	<.001***	-186.81	-170.55	99.40	-198.81	0.79	2	.675
	uncertain	8	-233.53	-211.85	124.76	-249.53	2.78	2	.249	-184.40	-162.72	100.20	-200.40	1.59	2	.451
ADL	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-205.06	-194.22	106.53	-213.06	2.16	1	.142	-194.16	-183.32	101.08	-202.16	4.15	1	.042*
	weak	6	-234.78	-218.52	123.39	-246.78	33.72	2	<.001***	-191.10	-174.85	101.55	-203.10	0.94	2	.625
	uncertain	8	-233.17	-211.49	124.58	-249.17	2.39	2	.302	-187.20	-165.52	101.60	-203.20	0.10	2	.953
LAM	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-204.54	-193.70	106.27	-212.54	1.64	1	.200	-195.16	-184.32	101.58	-203.16	5.14	1	.023*
	weak	6	-234.90	-218.65	123.45	-246.90	34.36	2	<.001***	-191.90	-175.64	101.95	-203.90	0.73	2	.693
	uncertain	8	-232.74	-211.06	124.37	-248.74	1.84	2	.399	-187.90	-166.22	101.95	-203.90	0.00	2	.998
TT	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-206.26	-195.42	107.13	-214.26	3.36	1	.067.	-194.65	-183.81	101.32	-202.65	4.63	1	.031*
	weak	6	-235.05	-218.79	123.52	-247.05	32.79	2	<.001***	-191.76	-175.51	101.88	-203.76	1.12	2	.572
	uncertain	8	-233.64	-211.96	124.82	-249.64	2.59	2	.274	-187.88	-166.21	101.94	-203.88	0.12	2	.943
Regularity	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-207.19	-196.35	107.59	-215.19	4.29	1	.038*	-195.04	-184.20	101.52	-203.04	5.02	1	0.025*
	weak	6	-235.33	-219.07	123.66	-247.33	32.14	2	<.000***	-192.21	-175.96	102.11	-204.21	1.17	2	0.557
	uncertain	8	-233.96	-212.28	124.98	-249.96	2.63	2	.268	-188.32	-166.64	102.16	-204.32	0.10	2	0.950
SampEn	null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
	strong	4	-203.23	-192.39	105.62	-211.23	0.33	1	.566	-190.84	-180.00	99.42	-198.84	0.82	1	.366
	weak	6	-234.66	-218.40	123.33	-246.66	35.43	2	<.001***	-187.61	-171.35	99.81	-199.61	0.78	2	.678
	uncertain	8	-234.94	-213.27	125.47	-250.94	4.28	2	.117	-184.07	-162.39	100.03	-200.07	0.46	2	.796

Note. Comprehension-reading time-ratio was used as dependent variable; models were compared against a null model containing only random effects. Regularity: determined first principal component. .p < .1, *p < .05, **p < .01, ***p < .001.

Hypothesis 3: Fixation-Based Measures as Predictors for Comprehension

Two mixed-effects models were set up including either only number of fixations per word or both, number of fixations per word and average fixation duration as fixed factors. Both models were again compared to a respective null model containing only random effects. Again, models were estimated separately for yes-/no-statements and wh-questions. The results are provided in **Table 8**. Based on model fit indices, the models that included only the number of fixations per word performed best. This effect was consistent across both comprehension item types, suggesting that there was no benefit in adding average fixation duration to the model. Our results replicate those of Southwell and colleagues (2020) insofar that the number of fixations indeed predicted text comprehension. However, we found the relationship between both measures to be inversed: better comprehension was predicted by less fixations.

Table 8

Model comparison: Fixation-based measures for comprehension prediction

Model		Yes-/no-statements							Wh-questions						
	nPar	AIC	BIC	logLik	deviance	χ^2	df	p	AIC	BIC	logLik	deviance	χ^2	df	p
null	3	-204.9	-196.77	105.45	-210.9				-192.02	-183.89	99.01	-198.02			
nFix	4	-281.73	-270.89	144.86	-289.73	78.82	1	<.001***	-197.95	-187.12	102.98	-205.95	7.94	1	.005**
nFix + FixDur	5	-283.9	-270.35	146.95	-293.9	4.18	1	.041 *	-195.96	-182.41	102.98	-205.96	0.00	1	.954

Note. Comprehension-reading time-ratio was used as dependent variable; models were compared against a null model containing only random effects. nFix: number of fixations; FixDur: average fixation duration. *p < .05, **p < .01, ***p < .001.

Exploration: Syntheses of Approaches

In an exploratory attempt to merge both approaches, either the best regularity-based predictor (RR), or the best fixation-based predictor (number of fixations per word), or both were included in linear mixed-effects models. Again, all models were computed separately for both comprehension item types, and tested against a null model containing only random

factors. Model comparisons are displayed in **Table 9** below. Compared to the null model, all other models performed better. The combined model showed the best model fit.

Table 9

Model comparison: Fixations and/or regularity-based measures for comprehension prediction

Model	*nPar*	Yes-/no-statements							Wh-questions						
		AIC	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*
null	3	-204.90	-196.77	105.45	-210.90				-192.02	-183.89	99.01	-198.02			
RR	4	-230.55	-219.71	119.28	-238.55	27.65	1	<.001 ***	-197.76	-186.93	102.88	-205.76	7.75	1	.005 **
nFix	4	-281.73	-270.89	144.86	-289.73	78.82	1	<.001 ***	-197.95	-187.12	102.98	-205.95	7.94	1	.005 **
RR+nFix	5	-283.96	-270.42	146.98	-293.96	83.06	2	<.001 ***	-198.03	-184.48	104.01	-208.03	10.01	2	.007 **

Note. Comprehension-reading time-ratio was used as dependent variable; models were compared against a null model containing only random effects. nFix: number of fixations. *p < .05, **p < .01, ***p < .001.

Discussion

The current study investigated the assumption of RTR that regularity measures (here RQA measures and SampEn based on gaze steps) capture the reading process, and are informative about text comprehension. Regularity measures successfully predicted comprehension-reading time scores, but not all the measures, and only RR and the reduced regularity component were consistent predictors for both, yes-/no-statements and wh-questions. Furthermore, the relation between regularity and comprehension ratios seemed to contradict the expected pattern of results: higher comprehension ratios entailed less regularity. We further tested the robustness of the effects of regularity measures in relation to comprehension. All regularity measures that successfully predicted comprehension-reading time-ratios for wh-question were in line with the assumption of strong invariance. Thus, effects of regularity measures turned out to be invariant across reading conditions. Turning towards yes-/no-statements, however, we saw that differences in comprehension ratios were mainly driven by the different reading conditions.

Somehow striking about the results are the differences between the two item types used to assess comprehension. Even though all items were previously rated, piloted, and selected based on a confirmatory factor analysis (cf. Tschense & Wallot, 2022b), they still seem to assess text comprehension differently, at least to some degree. Descriptively, participants' comprehension scores for yes-/no-statements were higher (M = 72.80%) than the achieved ones for wh-questions (M = 54.31%), suggesting an inherent difference in difficulty. Thus, wh-questions might reflect different degrees of text comprehension more precisely, whereas yes-/no-statements could still be answered sufficiently well with decreasing reading accuracy.

Furthermore, it has to be noted that regularity measures did not predict participants' "raw" comprehension scores, and analyses addressing the robustness yielded condition

effects only. In contrast, when adjusted for reading time, participants' comprehension was successfully be predicted by measures of regularity. These findings are in line with previous results by Wallot and colleagues (2014), where ratio scores revealed strongest and more consistent effects across reading tasks. A possible interpretation could be that comprehension as assessed by a post-hoc question battery alone is not a good-enough measure to depict the result of a complex and dynamic process as text reading. So it might be that dividing participants' comprehension scores by their reading time accounts for another component of the reading process, thus, resulting in a more informative measure.

Southwell and colleagues (2020) found that comprehension was predicted by more and longer fixations. While we also established the number of fixations to be predictive of comprehension, higher comprehension ratios were associated with less fixations. Comparing both studies, quite considerable differences are evident with regards to the study goal and actual tasks and/or instructions (tracking mind-wandering during reading vs. differently paced reading), the presented texts (non-fictional, expository texts vs. fictional, narrative texts), comprehension assessment (multiple-choice, surface-level items vs. surface and inference items and two item types), and later utilization thereof (accuracy scores vs. accuracy-reading time-ratios). Based on the currently available information, we can only speculate whether one or more of these factors may contribute to the divergent pattern of results.

Combining regularity- and fixation-based measures to predict participants' reading comprehension (descriptively) explained more variance than both measures individually did. However, what this added value actually constitutes is still unclear. It could be argued that fixations, as a more global and static measure of the reading process, might tell us more about task-, text-, and reader-driven factors. In contrast, regularity measures capturing the dynamics of the reading process might tap more into the interplay of these factors, such as a reader's

ability to flexibly adapt to situational changes. First pointers in this direction were found for writing, where recurrence measures indicated language proficiency as a reflection of effortfulness during text production (Haake et al., 2022). But more research is needed in order to disentangle commonalities and differences between these measures.

The current study bears some limitations which should be considered for interpretation. First, it has to be noted that the manipulation of participants' reading comprehension was only partly successful: While we saw differences for normal and slow reading compared to fast reading, comprehension between normal and slow reading did not differ. Also, the resulting accuracy-reading time-ratios suggested that participants' comprehension did not drop in a comparable rate as reading speed increased. Furthermore, we cannot rule out that the different reading instructions solely affected comprehension, and did not evoke certain coping mechanisms such as switching reading strategies. Another way to achieve the intended manipulation would be to use texts of different levels of difficulty, which would reduce the compliance of the reader and the instructor.

In conclusion, we demonstrated RTR to be a promising tool to capture the reading process and to predict text comprehension. However, more research is needed to further investigate how specific task demands, and readers' cognitive abilities and reading strategies are reflected within this framework.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) by grants to Sebastian Wallot (project numbers 397523278 and 442405852). We would like to thank Maria Raab, Franziska Roth and Nadejda Rubinskii for their help with stimulus selection, data collection and preprocessing.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4 (R package version 1.1-30). *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blanchard, H. E., & Iran-Nejad, A. (1987). Comprehension processes and eye movement patterns in the reading of surprise-ending stories. *Discourse Processes, 10*(1), 127–138. https://doi.org/10.1080/01638538709544663

Booth, C. R., Brown, H. L., Eason, E. G., Wallot, S., & Kelty-Stephen, D. G. (2018). Expectations on hierarchical scales of discourse: Multifractality predicts both short- and long-range effects of violating gender expectations in text reading. *Discourse Processes, 55*(1), 12–30. https://doi.org/10.1080/0163853X.2016.1197811

Breznitz, Z., Shaul, S., Horowitz-Kraus, T., Sela, I., Nevat, M., & Karni, A. (2013). Enhanced reading by training with imposed time constraint in typical and dyslexic adults. *Nature Communications, 4*, 1486. https://doi.org/10.1038/ncomms2488

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84–95. http://www.jstor.org/stable/40016440

Coco, M. I., Mønster, D., Leonardi, G., Dale, R., & Wallot, S. (2021). Unidimensional and multidimensional methods for recurrence quantification analysis with CRQA (R package version 2.0.3), *The R Journal, 13*(1), 145–163. https://doi.org/10.32614/RJ-2021-062

Dalmaijer, E.S., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments. *Behavior Research Methods, 46*, 913–921. https://doi.org/10.3758/s13428-013-0422-2

Dyson, M. and Haselgrove, M. (2000), The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of Research in Reading, 23*(2), 210–223. https://doi.org/10.1111/1467-9817.00115

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*(4), 777–813. https://doi.org/10.1037/0033-295x.112.4.777

Engbert, R., Sinn, P., Mergenthaler, K., & Trukenbrod, H. (2015). *Microsaccade toolbox for R*. Potsdam Mind Research Repository. https://t1p.de/ochq

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*(5), 263–279. https://doi.org/10.1017/S0140525X11001841

Graesser, A. C., Millis, K. K., and Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163–189. https://doi.org/10.1146/annurev.psych.48.1.163

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193–202. https://doi.org/10.3758/BF03195564

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*(3), 518–565. https://doi.org/10.1037/0033-295X.103.3.518

Haake, L., Wallot, S., Tschense, M., & Grabowski, J. (2022). Global temporal typing patterns in foreign language writing: exploring language proficiency through recurrence quantification analysis (RQA). *Reading and Writing,* 1–33. https://doi.org/10.1007/s11145-022-10331-0

Heister, J., Würzner, K. M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau, 62*(1), 10-20. https://doi.org/10.1026/0033-3042/a000029

Holden, J. G., & Van Orden, G. (2002). Reading. In M. A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks* (2nd ed., pp. 951–955). MIT Press.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of Individual Differences in Reading Comprehension and Reading Fluency. *Journal of Educational Psychology, 95*(4), 719–729. https://doi.org/10.1037/0022-0663.95.4.719

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kapteijns, B., & Hintz, F. (2021). Comparing predictors of sentence self-paced reading times: Syntactic complexity versus transitional probability metrics. *PLoS ONE, 16*(7), e0254546. https://doi.org/10.1371/journal.pone.0254546

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5*(3), 257–274. https://doi.org/10.1016/0010-0285(73)90036-4

Kuznetsov, N., Bonnette, S., & Riley, M.A. (2013). Nonlinear time series methods for analyzing behavioural sequences. In K. Davids, R. Hristovski, D. Araujo, N. Balague Serre, C. Button, & P. Passos (eds.), *Complex Systems in Sport* (pp. 85–104). Routledge.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models (R package version 3.1-3). *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

LeVasseur, V. M., Macaruso, P., Palumbo, L. C., & Shankweiler, D. (2006). Syntactically cued text facilitates oral reading fluency in developing readers. *Applied Psycholinguistics, 27*(3), 423–445. https://doi.org/10.1017/S0142716406060346

LeVasseur, V. M., Macaruso, P., & Shankweiler, D. (2008). Promotion gains in reading fluency: A comparison of three approaches. *Reading and Writing, 21*, 205–230. https://doi.org/10.1007/s11145-007-9070-1

MathWorks Inc. (2020). *MATLAB* (version R2020b). The MathWorks Inc. https://www.mathworks.com

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports, 438*(5–6), 237–329. https://doi.org/10.1016/j.physrep.2006.11.001

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*, 314–324. https://doi.org/10.3758/s13428-011-0168-7

Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition, 30*, 551–561. https://doi.org/10.3758/BF03194956

Mézière, D. C., Yu, L., McArthur, G., Reichle, E. D., & von der Malsburg, T. (2023a). Scanpath regularity as an index of reading comprehension. *Scientific Studies of Reading*, 1–22. https://doi.org/10.1080/10888438.2023.2232063

Mézière, D. C., Yu, L., Reichle, E. D., Von Der Malsburg, T., & McArthur, G. (2023b). Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly, 58*, 425–449. https://doi.org/10.1002/rrq.498

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*, 45–52. https://doi.org/10.3758/BF03193811

O'Brien, B. A., & Wallot, S. (2016). Silent reading fluency and comprehension in bilingual children. *Frontiers in Psychology, 7*, 1265. https://doi.org/10.3389/fpsyg.2016.01265

O'Brien, B. A., Wallot, S., Haussmann, A., & Kloos, H. (2014). Using complexity metrics to assess silent reading fluency: A cross-sectional study comparing oral and silent reading. *Scientific Studies of Reading, 18*(4), 235–254. https://doi.org/10.1080/10888438.2013.862248

Perfetti, C. A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67*(4), 461–469. https://doi.org/10.1037/h0077013

R Core Team (2022). *R: A language and environment for statistical computing* (version 4.2.0). R Foundation for Statistical Computing. https://www.R-project.org/

Ramdani, S., Seigle, B., Lagarde, J., Bouchara, F., & Bernard, P. L. (2009). On the use of sample entropy to analyze human postural sway data. *Medical Engineering & Physics, 31*(8), 1023–1031. https://doi.org/10.1016/j.medengphy.2009.06.004

Rayner, K. (1985). Do faulty eye movements cause dyslexia? *Developmental Neuropsychology, 1*(1), 3–15. https://doi.org/10.1080/87565648509540294

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

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. https://doi.org/10.1207/s1532799xssr1003_3

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*(1), 1–21. https://doi.org/10.3758/PBR.16.1.1

Revelle, R. (2023). *psych: Procedures for psychological, psychometric, and personality research* (R package version 2.3.3). Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psychn

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology–Heart and Circulatory Physiology, 278*(6), H2039–H2049. https://doi.org/10.1152/ajpheart.2000.278.6.H2039

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition, 131*(1), 1–27. https://doi.org/10.1016/j.cognition.2013.11.018

Schroeder, S. (2011). What readers have and do: Effects of students' verbal ability and reading time components on comprehension with and without text availability. *Journal of Educational Psychology, 103*(4), 877–896. https://doi.org/10.1037/a0023731

Southwell, R., Gregg, J., Bixler, R., & D'Mello, S. K. (2020). What eye movements reveal about later comprehension of long connected texts. *Cognitive Science, 44*(10), e12905. https://doi.org/10.1111/cogs.12905

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition, 116*(1), 71–86. https://doi.org/10.1016/j.cognition.2010.04.002

Stephen, D. G., & Mirman, D. (2010). Interactions dominate the dynamics of visual cognition. *Cognition, 115*(1), 154–165. https://doi.org/10.1016/j.cognition.2009.12.010

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition, 105*(2), 477–488. https://doi.org/10.1016/j.cognition.2006.10.009

Teng, D. W., Wallot, S., & Kelty-Stephen, D. G. (2016). Single-word recognition need not depend on single-word features: Narrative coherence counteracts effects of single-word features that lexical decision emphasizes. *Journal of Psycholinguistic Research, 45*(6), 1451–1472. https://doi.org/10.1007/s10936-016-9416-4

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual differences in eye-movements during reading: Working memory and speed-of-processing effects. *Journal of Eye Movement Research, 5*(1), 1–16. https://doi.org/10.16910/jemr.5.1.5

Tschense, M., & Wallot, S. (2022a). Using measures of reading time regularity (RTR) to quantify eye movement dynamics and how they are shaped by linguistic information. *Journal of Vision, 22*(6), 9. https://doi.org/10.1167/jov.22.6.9

Tschense, M., & Wallot, S. (2022b). Modeling items for text comprehension assessment using confirmatory factor analysis. *Frontiers in Psychology, 13*, 966347. https://doi.org/10.3389/fpsyg.2022.966347

Velan, H., & Frost, R. (2007). Cambridge University versus Hebrew University: The impact of letter transposition on reading English and Hebrew. *Psychonomic Bulletin & Review, 14*(5), 913–918. https://doi.org/10.3758/BF03194121

Wallot, S. (2014). From "cracking the orthographic code" to "playing with language": Toward a usage-based foundation of the reading process. *Frontiers in Psychology, 5*, 891. https://doi.org/10.3389/fpsyg.2014.00891

Wallot, S. (2016). Understanding reading as a form of language-use: A language game hypothesis. *New Ideas in Psychology, 42*, 21–28. https://doi.org/10.1016/j.newideapsych.2015.07.006

Wallot, S. (2017). Recurrence quantification analysis of processes and products of discourse: A tutorial in R. *Discourse Processes, 54*(5–6), 382–405. https://doi.org/10.1080/0163853X.2017.1297921

Wallot, S., Hollis, G., & van Rooij, M. (2013). Connected text reading and differences in text reading fluency in adult readers. *PLoS ONE, 8*(8), e71914. https://doi.org/10.1371/journal.pone.0071914

Wallot, S., O'Brien, B., Coey, C. A., & Kelty-Stephen, D. (2015). Power-law fluctuations in eye movements predict text comprehension during connected text reading. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2583–2588). Cognitive Science Society.

Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, H., & Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1745–1765. https://doi.org/10.1037/xlm0000030

Xiong, J., Yu, L., Veldre, A., Reichle, E. D., & Andrews, S. (2023). A multitask comparison of word- and character-frequency effects in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(5), 793–811. https://doi.org/10.1037/xlm0001192

Zbilut, J. P., & Webber, C. L., Jr. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A, 171*(3–4), 199–203. https://doi.org/10.1016/0375-9601(92)90426-M

Ziegler, J. C., Perry, C., & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology, 12*(3), 413–430. https://doi.org/10.1080/09541440050114570

Zipke, M. (2007). The role of metalinguistic awareness in the reading comprehension of sixth and seventh graders. *Reading Psychology, 28*(4), 375–396. https://doi.org/10.1080/02702710701260615

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 386–397. https://doi.org/10.1037/0278-7393.21.2.386

Supplemental Materials

Supplement 1: Items for Comprehension Assessment

Short story 1

Code	Level	Item	Correct answer
AU01_13	micro	Da der Zug morgens dreckig und überfüllt war, stellten sich einige Schulkinder außen auf die Trittbretter des Zuges.	yes
AU01_15	micro	Der Großbauer bezahlte fünf D-Mark für 50 kg gepflückte Erbsen.	no
AU01_24	micro	Bei Tisch saß immer der älteste Lehrling des Rittergutes neben dem Gutsherrn.	no
AU01_26	micro	Freiherr von Waldern-Biegnitz war ein Heimatvertriebener, der nun in einem Forschungsinstitut arbeitete.	yes
AU01_28	micro	Der Höhepunkt des ersten Ausbildungsjahres war das Erntedankfest, welches üppig gefeiert wurde.	yes
AU01_33	micro	Die vorherigen Auszubildenden des Hofguts waren eines Nachts vom Hofgut weggelaufen.	yes
AU01_47	inference	Die Protagonistin durfte sich ihre Ausbildung frei aussuchen.	no
AU01_52	inference	Das Erntedankfest, das als Belohnung für die Bediensteten des Rittergutes gedacht war, führte häufig zu Reibereien unter den Arbeitern.	no
AU01_53	inference	Freiherr von Waldern-Biegnitz hatte sich in die Protagonistin verliebt, was jedoch gegen geltende Etikette verstoß.	yes
AU01_54	inference	Freiherr von Waldern-Biegnitz setzte sich über die damalige Etikette hinweg und gestand der Protagonistin offen seine Liebe.	no
WF01_01	micro	Warum fand der Unterricht damals zum Teil erst am Nachmittag statt?	open-ended
WF01_03	micro	Warum stellten sich einige Schulkinder morgens auf die Trittbretter des Zuges?	open-ended
WF01_05	micro	Wie begrüßte Onkel Karl die Protagonistin und ihre Schwester immer?	open-ended
WF01_07	micro	Welche Spitznamen gaben die Arbeiter der Gutsherrschaft?	open-ended
WF01_13	inference	Welches Ereignis auf dem Rittergut stellte eine Ausnahme zur sonst so strengen Etikette dar?	open-ended
WF01_14	inference	Warum nahm Freiherr von Waldern-Biegnitz nicht am Erntedankfest teil?	open-ended
WF01_15	inference	Welches ihrer beiden Lehrjahre gefiel der Protagonistin besser?	open-ended

Short story 2

Code	Level	Item	Correct answer
AU02_07	micro	Die Stadt, die Jack suchte, sollte sich inmitten der Wüste befinden.	yes
AU02_09	micro	Jack bekam mehrere Auszeichnungen, nachdem er im Zweiten Weltkrieg gegen Deutschland und Japan kämpfte.	yes
AU02_11	micro	Jack empfand es als Ehre, einer der beiden Piloten zu sein, welche die Atombomben abwarfen.	yes
AU02_17	micro	Nach der Trennung von Mary zog Jack in eine kleine Wohnung in der Nachbarstadt.	no
AU02_27	micro	Mit seinen neuen Arbeitskollegen spielte Jack gern Karten oder schaute Baseball.	yes
AU02_28	micro	Als Patricias Tasche riss, half Jack ihr beim Auflesen ihrer Einkäufe und schenkte ihr seine eigene Tasche.	yes
AU02_31	micro	Um Patricia wiederzusehen, aß Jack fast jeden Tag in dem Restaurant, in dem sie arbeitete.	yes
AU02_38	micro	Nachdem er eines Morgens inmitten des Sandes der Halbwüste aufwachte, irrte Jack orientierungslos umher, bis ihn seine Kräfte verließen.	yes

AU02_43	inference	Als er einwilligte, die Atombombe über Japan abzuwerfen, war Jack die Tragweite seines Handelns völlig bewusst.	no
AU02_46	inference	Die Unzufriedenheit darüber, keine Kinder haben zu können, führte zur Scheidung von Jack und Mary.	no
AU02_50	inference	Jack war nur dann glücklich, wenn er erfolgreich einer bedeutungsvollen Arbeit nachging und das Gefühl hatte, gebraucht zu werden.	yes
AU02_57	inference	Jack suchte vergebens nach seiner nicht existierenden Familie, nachdem er aus dem Koma erwachte.	no
WF02_01	micro	Welche Stadt hatte der Erzähler auf seiner Reise bereits besucht?	open-ended
WF02_02	micro	Wonach suchte der Mann, den der Erzähler in der Wüste traf?	open-ended
WF02_03	micro	Wo absolvierte Jack seine Ausbildung?	open-ended
WF02_05	micro	Warum zog Jack nach der Trennung von seiner ersten Frau in ein Motelzimmer?	open-ended
WF02_07	micro	Woran arbeiteten Jack und seine Kollegen in der Flugzeugfirma?	open-ended
WF02_10	micro	Warum stellten Jack und Patricia eine Haushaltshilfe ein?	open-ended
WF02_12	inference	Warum übernahm Jack das Abwerfen der Atombombe?	open-ended
WF02_13	inference	Warum trennten sich Jack und seine erste Frau?	open-ended
WF02_15	inference	Warum hatten es weder Jack noch Patricia eilig, eine Beziehung einzugehen?	open-ended
WF02_16	inference	Was realisierte Jack gerade noch, bevor er in der Realität erwachte?	open-ended

Short story 3

Code	Level	Item	Correct answer
AU03_07	micro	Ritas Mann beeilte sich nach der Arbeit, um pünktlich zum Abendessen Zuhause zu sein.	no
AU03_10	micro	Als ihr Mann von der Arbeit nach Hause kam, verhielt sich Rita misstrauisch und distanziert.	yes
AU03_14	micro	Rita war von dem technischen Fortschritt heutzutage überfordert.	yes
AU03_15	micro	Rita erwog die Möglichkeit, dass ihr Mann selbst etwas mit dem Täuschungsmanöver zu tun hatte.	yes
AU03_26	micro	Rita wollte ihren Mann im Krankenhaus untersuchen lassen, um seine Nichtmenschlichkeit beweisen zu können.	yes
AU03_27	micro	Im grellen Licht des Krankenhauses fühlte sich Rita unwohl.	yes
AU03_37	micro	Der Psychiater beschrieb Rita dem Oberarzt als ängslich-depressiv und psychomotorisch unruhig.	yes
AU03_41	inference	Die Notiz, die Rita fand, wurde vermutlich von jemand anderem als ihrem Mann verfasst.	yes
AU03_45	inference	Ritas Mann war der Alleinverdiener der Familie.	yes
AU03_47	inference	Neben ihrem primären Verdacht zog Rita auch in Erwägung, dass ihr Mann sich in eine andere Frau verliebt hatte.	no
AU03_51	inference	Rita hatte nur noch sehr wenige Freunde, denen sie sich anvertrauen konnte.	yes
AU03_57	inference	Ritas Erkrankung ist vermutlich genetisch veranlagt.	yes
WF03_01	micro	Was stand auf der Notiz, die Rita gefunden hat?	open-ended
WF03_05	micro	Wie interpretierte Rita, dass ihr Mann die Kapern ohne Protest gegessen hatte?	open-ended
WF03_06	micro	Wie schätzte der Polizist Ritas Zustand ein?	open-ended
WF03_10	micro	Was verschwieg Rita dem Psychiater bei der Anamnese?	open-ended
WF03_11	inference	Warum nahm Rita ein Foto ihres Mannes aus dem Fotoalbum heraus?	open-ended
WF03_13	inference	Warum zeigte Ritas Mann keinerlei Reaktion auf Ritas provokatives Verhalten während des Abendessens?	open-ended
WF03_14	inference	Warum wandte sich Rita an die Polizei?	open-ended

Supplement 2: Results of principal component analysis (PCA)

Table 10

Standardized loadings (pattern matrix) based upon correlation matrix

	PC1	PC2	PC3	PC4	PC5	PC6	h2	u2	com
RR	0.80	-0.15	0.58	0.09	0.00	0.00	1.00	0.00	1.90
DET	0.90	0.43	-0.05	0.02	-0.05	0.01	1.00	0.00	1.40
MDL	0.85	-0.37	-0.30	0.23	0.00	0.00	1.00	0.00	1.80
ADL	0.96	-0.22	-0.09	-0.18	0.02	0.02	1.00	0.00	1.20
LAM	0.84	0.53	-0.05	0.04	0.04	-0.01	1.00	0.00	1.70
TT	0.97	-0.20	-0.03	-0.15	-0.01	-0.03	1.00	0.00	1.10

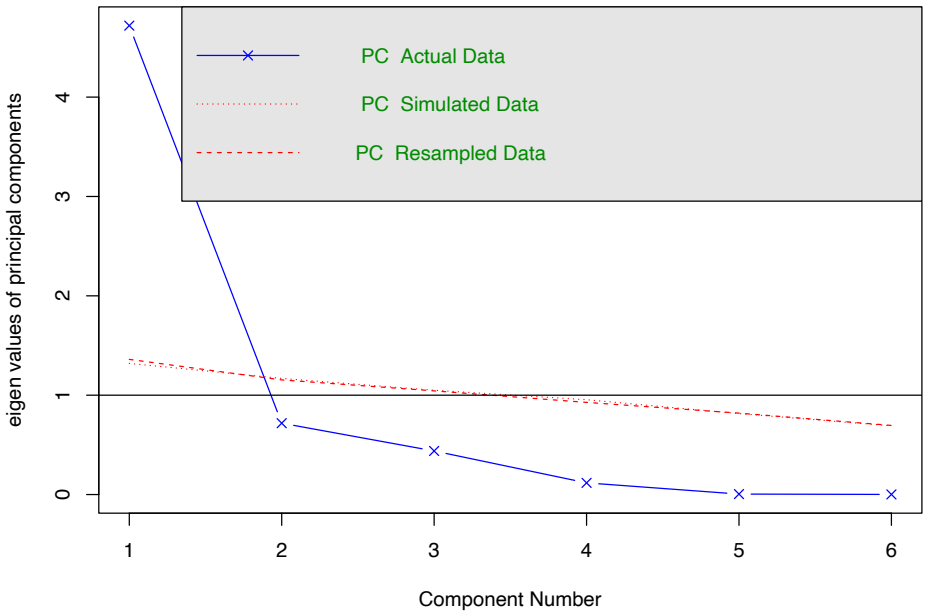
Table 11

Variance accounted for by principal components

	PC1	PC2	PC3	PC4	PC5	PC6
SS loadings	4.72	0.72	0.44	0.12	0.00	0.00
Proportion Var	0.79	0.12	0.07	0.02	0.00	0.00
Cumulative Var	0.79	0.91	0.98	1.00	1.00	1.00
Proportion Explained	0.79	0.12	0.07	0.02	0.00	0.00
Cumulative Proportion	0.79	0.91	0.98	1.00	1.00	1.00

Figure 1

Scree plots of parallel analysis



Supplement 3: Results for raw comprehension scores

Hypothesis 2: Strong vs. weak vs. uncertain invariance of effects

To further investigate the relationship between comprehension and regularity, three mixed-effects models were set up reflecting a strong, weak and uncertain invariance of effects according to our second hypothesis. All models were estimated separately per regularity measure and comprehension item type. The models were compared to a null model containing only the random-effects structure. Results are provided in **Table 12**. Regardless of the type of comprehension items, all regularity measures consistently provided support for the assumption of weak invariance of effects. All effects were driven by the additional fixed factor of reading condition.

Hypothesis 3: Fixation-based measures as predictors for comprehension

Two mixed-effects models were set up including either only number of fixations per word or both, number of fixations per word and average fixation duration as fixed factors. Both models were again compared to a respective null model containing only random effects. Again, models were estimated separately for yes-/no-statements and wh-questions. The results are provided in **Table 13**. Based on model fit indices, the models that included only the number of fixations per word performed best. This effect was consistent across both comprehension item types, suggesting that there was no benefit in adding average fixation duration to the model. In principle, these results replicate earlier findings by Southwell and colleagues (2020).

Exploration: Synthesis of approaches

In an attempt to merge both, the regularity- and fixation-based approach, linear mixed-effects models including number of fixations per word and/or the determined regularity component. Models were computed separately for both comprehension item types, and tested against a null model with only random factors (**Table 14**).

Table 12

Model comparison: Invariance of effects

Measure	Model		Yes-/no-statements							Wh-questions						
		nPar	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*
RR	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	913.12	923.96	-452.56	905.12	1.89	1	.169	1010.61	1021.45	-501.31	1002.61	2.64	1	.104
	weak	6	905.01	921.26	-446.5	893.01	12.12	2	.002**	981.03	997.29	-484.52	969.03	33.58	2	<.001***
	uncertain	8	908.37	930.05	-446.19	892.37	0.63	2	.728	984.63	1006.31	-484.32	968.63	0.40	2	.817
DET	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	914.95	925.79	-453.47	906.95	0.06	1	.800	1011.23	1022.07	-501.62	1003.23	2.02	1	.156
	weak	6	904.68	920.93	-446.34	892.68	14.27	2	<.001***	977.97	994.22	-482.98	965.97	37.27	2	<.001***
	uncertain	8	907.95	929.62	-445.97	891.95	0.73	2	.694	981.83	1003.5	-482.91	965.83	0.14	2	.932
MDL	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	914.96	925.80	-453.48	906.96	0.05	1	.820	1013.11	1023.95	-502.55	1005.11	0.14	1	.706
	weak	6	904.97	921.22	-446.48	892.97	13.99	2	<.001***	982.86	999.12	-485.43	970.86	34.24	2	<.001***
	uncertain	8	906.80	928.48	-445.40	890.80	2.17	2	.339	986.29	1007.97	-485.15	970.29	0.57	2	.751
ADL	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	915.00	925.84	-453.50	907.00	0.01	1	.915	1013.21	1024.05	-502.60	1005.21	0.04	1	.833
	weak	6	904.47	920.72	-446.23	892.47	14.53	2	<.001***	980.90	997.16	-484.45	968.90	36.30	2	<.001***
	uncertain	8	908.17	929.85	-446.09	892.17	0.29	2	.864	984.01	1005.69	-484.01	968.01	0.89	2	.642
LAM	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	914.88	925.72	-453.44	906.88	0.13	1	.716	1010.96	1021.80	-501.48	1002.96	2.29	1	.130
	weak	6	904.72	920.98	-446.36	892.72	14.16	2	<.001***	978.87	995.12	-483.43	966.87	36.09	2	<.001***
	uncertain	8	907.89	929.56	-445.94	891.89	0.83	2	.659	982.77	1004.44	-483.38	966.77	0.10	2	.952
TT	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	915.00	925.84	-453.50	907.00	0.01	1	.909	1013.25	1024.09	-502.63	1005.25	0.00	1	.979
	weak	6	904.62	920.88	-446.31	892.62	14.38	2	<.001***	980.76	997.02	-484.38	968.76	36.49	2	<.001***
	uncertain	8	908.28	929.96	-446.14	892.28	0.34	2	.844	984.02	1005.69	-484.01	968.02	0.75	2	.689
Regularity	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	914.93	925.77	-453.47	906.93	0.08	1	.777	1013.24	1024.08	-502.62	1005.24	0.01	1	.934
	weak	6	904.81	921.07	-446.40	892.81	14.12	2	<.001***	980.29	996.55	-484.15	968.29	36.95	2	<.001***
	uncertain	8	907.87	929.55	-445.94	891.87	0.93	2	.627	983.95	1005.63	-483.97	967.95	0.34	2	.842
SampEn	null	3	913.01	921.14	-453.51	907.01				1011.25	1019.38	-502.63	1005.25			
	strong	4	914.22	925.06	-453.11	906.22	0.79	1	.374	1012.74	1023.58	-502.37	1004.74	0.51	1	.473
	weak	6	904.98	921.24	-446.49	892.98	13.24	2	.001**	982.80	999.06	-485.40	970.80	33.93	2	<.001***
	uncertain	8	907.57	929.25	-445.79	891.57	1.41	2	.494	986.77	1008.45	-485.38	970.77	0.03	2	.983

Note. Comprehension score was used as dependent variable; models were compared against a null model containing only random effects. Regularity: determined first principal component. .$p < .1$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table 13

Model comparison: Fixation-based measures for comprehension prediction

Model		Yes-/no-statements							Wh-questions						
	nPar	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*
null	3	913.01	921.14	-453.51	907.01				1011.25	1019.40	-502.63	1005.25			
nFix	4	900.28	911.12	-446.14	892.28	14.73	1	<.001 ***	998.52	1009.40	-495.26	990.52	14.73	1	<.001 ***
nFix + FixDur	5	900.29	913.84	-445.15	890.29	1.99	1	.158	997.41	1011.00	-493.70	987.41	3.12	1	.077 .

Note. Comprehension score was used as dependent variable; models were compared against a null model containing only random effects. nFix: number of fixations; FixDur: average fixation duration. .p < .1, *p < .05, **p < .01, ***p < .001.

Table 14

Model comparison: Fixation- and/or regularity-based measures for comprehension prediction

Model		Yes-/no-statements							Wh-questions						
	nPar	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*	*AIC*	*BIC*	*logLik*	*deviance*	χ^2	*df*	*p*
null	3	913.01	921.14	-453.51	907.01				1011.25	1019.40	-502.63	1005.25			
nFix	4	900.28	911.12	-446.14	892.28	14.73	1	<.001***	998.52	1009.40	-495.26	990.52	14.73	1	<.001***
Regularity	4	914.93	925.77	-453.47	906.93	0.00	0	1.000	1013.24	1024.10	-502.62	1005.24	0.00	0	1.000
nFix + Regularity	5	902.28	915.83	-446.14	892.28	14.65	1	<.001***	1000.27	1013.80	-495.14	990.27	14.97	1	<.001***

Note. Comprehension score was used as dependent variable; models were compared against a null model containing only random effects. nFix: number of fixations; Regularity: determined first principal component. .p < .1, *p < .05, **p < .01, ***p < .001.

APPENDIX: CURRICULUM VITAE

Monika Tschense

✉ Leuphana University of Lüneburg
Institute for Sustainability Education and Psychology
Universitätsallee 1
21335 Lüneburg, Germany

@ monika.tschense@leuphana.de

+49 (0) 4131 677 1671

Academic Appointments

Since 01/2021	**Researcher**, Institute for Sustainability Education and Psychology, Leuphana University of Lüneburg, Germany
Since 01/2021	**Guest Researcher**, Research Group for Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Frankfurt/M., Germany
11/2018 - 12/2020	**Researcher**, Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Frankfurt/M., Germany
05/2017 - 07/2018	**Researcher**, Research Group for Clinical Linguistics, Institute of German Linguistics, Philipps University of Marburg, Germany
06/2016 - 05/2017	**Research Assistant**, Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Frankfurt/M., Germany
04/2013 - 08/2014	**Student Assistant**, Department of Psycholinguistics and Teaching of German, Goethe University Frankfurt/M., Germany

Education

Since 11/2018	**PhD Candidate in Psychology**, Leuphana University of Lüneburg & Max Planck Institute for Empirical Aesthetics, Frankfurt/M., Germany
10/2015 - 04/2019	**Master of Arts in Linguistics: Cognition and Communication**, Philipps University of Marburg, Germany
10/2014 - 03/2017	**Master of Arts in Clinical Linguistics**, Philipps University of Marburg, Germany
10/2011 - 09/2014	**Bachelor of Arts in Linguistics**, Goethe University Frankfurt/M., Germany
06/2011	**Abitur**

Scholarships

10/2016 - 09/2017 **Deutschlandstipendium**, Philipps University of Marburg, Germany & Federal Government of Germany

Publications

Tschense, M., & Wallot, S. (2023). *Using recurrence quantification analysis to measure reading comprehension for long, connected texts* [Manuscript submitted for publication]. Institute for Sustainability Psychology, Leuphana University of Lüneburg.

Wallot, S., Irmer, J.P., **Tschense, M.**, Kuznetsov, N., Højlund, A. and Dietz, M. (2023). A multivariate method for dynamic system analysis: Multivariate detrended fluctuation analysis using generalized variance. *Topics in Cognitive Science*. Advance online publication. https://doi.org/10.1111/tops.12688

Haake, L., Wallot, S., **Tschense, M.**, & Grabowski, J. (2022). Global temporal typing patterns in foreign language writing: exploring language proficiency through recurrence quantification analysis (RQA). *Reading and Writing*, 1–33. https://doi.org/10.1007/s11145-022-10331-0

Tschense, M., & Wallot, S. (2022b). Modeling items for text comprehension assessment using confirmatory factor analysis. *Frontiers in Psychology, 13*, 966347. https://doi.org/10.3389/fpsyg.2022.966347

Tschense, M., & Wallot, S. (2022a). Using measures of reading time regularity (RTR) to quantify eye movement dynamics and how they are shaped by linguistic information. *Journal of Vision, 22*(6), 9, 1-21. https://doi.org/10.1167/jov.22.6.9

Vesker, M., Bahn, D., Kauschke, C., **Tschense, M.**, Degé, F., & Schwarzer, G. (2018). Auditory emotion word primes influence emotional face categorization in children and adults, but not vice versa. *Frontiers in Psychology, 9*, 618. https://doi.org/10.3389/fpsyg.2018.00618

Talks

Tschense, M., & Wallot, S. (2023). Regularity in Eye Movement Data During Text Reading. *International Symposium on Recurrence Plots*. University of Tsukuba, Ibaraki, Japan.

Tschense, M., & Wallot, S. (2023). Reading Time Regularity. *Workshop on Nonlinear Methods for Psychological and Social Science (NLM 2023)*. Leuphana University of Lüneburg, Germany.

Tschense, M., & Wallot, S. (2022). Nonlinear Dynamics of Eye Movements during Text Reading. *9th International Conference of the German Cognitive Linguistics Association (DGKL/GCLA-9)*. University of Erfurt, Germany (virtual).

Tschense, M., & Wallot, S. (2021). Disentangling Text Comprehension. *31st Annual Meeting of the Society for Text & Discourse (ST&D)*. Chicago, Illinois, USA (virtual).

Tschense, M., & Wallot, S. (2021). Looking in patterns: Recurrence quantification analysis (RQA) of eye movements. *63rd Conference of Experimental Psychologists (TeaP)*. Ulm University, Germany (virtual).

Wallot, S., & Tschense, M. (2021). Multidimensional Detrended Fluctuation Analysis (MdDFA) for the quantification of global long-memory processes and its application to EEG data. *63rd Conference of Experimental Psychologists (TeaP)*. Ulm University, Germany (virtual).

Tschense, M. (2021). Nonlinear Dynamics of Reading and Text Comprehension. *Research Colloquium*. Institute for Psychology, Leuphana University of Lüneburg, Germany.

Tschense, M., & Wallot, S. (2020). Nonlinear Dynamics of Eye Movements during Text Reading. *9th International Conference of the German Cognitive Linguistics Association (DGKL/GCLA-9)*. University of Erfurt, Germany. Canceled due to COVID-19.

Tschense, M., & Wallot, S. (2020). Nonlinear dynamics of text reading: Recurrence quantification analysis of eye movements. *30th Annual Meeting of the Society for Text & Discourse (ST&D)*. Atlanta, Georgia, USA (virtual).

Tschense, M., & Wallot, S. (2020). Looking in patterns: Recurrence quantification analysis (RQA) of eye movements. *62nd Conference of Experimental Psychologists (TeaP)*. University of Jena, Germany. Canceled due to COVID-19.

Wallot, S., & Tschense, M. (2020). Multidimensional Detrended Fluctuation Analysis (MdDFA) for the quantification of global long-memory processes and its application to EEG data. *62nd Conference of Experimental Psychologists (TeaP)*. University of Jena, Germany. Canceled due to COVID-19.

Posters

Tschense, M., & Wallot, S. (2022). *Text Comprehension: Mental Representation and Assessment*. 15th Biannual Conference of the German Society for Cognitive Science (KogWis 2022). University of Freiburg, Germany.

Tschense, M., Domahs, U., & Kauschke, C. (2018). *Behavioral and Electrophysiological Correlates of Emotion Term Processing*. Multiple approaches to the perception of speech (MAPS). University of Groningen, Netherlands.

Kauschke, C., Schwarzer, G., Vesker, M., Tschense, M., & Bahn, D. (2017). *The interplay of verbal and facial information in emotion categorization*. On-site Review of the Collaborative Research Centre “Cardinal Mechanisms of Perception”. University of Marburg, Germany.