Measuring Suspense in Real Time: A New Experimental Methodology

MARIA BENTZ [iD]

MAYA CORTEZ ESPINOZA [iD]

VESELA SIMEONOVA [iD]

TILMANN KÖPPE [iD]

EDGAR ONEA [iD]

*Author affiliations can be found in the back matter of this article

ABSTRACT

Narrative suspense represents a multifaceted cognitive-emotional response triggered by unfolding storylines. Recognized as a critical aesthetic element in literary narratives, its investigation has gained momentum across psychology, narratology, and related disciplines, with a focus on its textual, emotional, and cognitive underpinnings. The empirical foundations of such studies hinge critically on our capacity to quantify suspense, establishing a dependable suspense arc for narratives. Despite a plethora of methods available within current literature, all with their respective merits and limitations, the empirical measurement of suspense, particularly for written texts, poses a substantial challenge. This paper introduces a novel experimental paradigm to construct a high-resolution suspense arc for extended literary narratives within a laboratory setting. The methodology is exemplified using a German translation of Arthur Conan Doyle's *The Brazilian Cat*.

CORRESPONDING AUTHOR:
Maria Bentz

University of Göttingen, DE

maria.bentz@uni-goettingen.de

INTRODUCTION

Narrative suspense is a fundamental feature of the literary reception that encompasses a range of emotional and cognitive responses to a written text (taken here as a metaphor for any sort of medial presentation of a narrative plot). Broadly speaking, suspense is responsible for keeping readers engaged with the text and is closely related to a range of emotional and cognitive responses such as story world absorption, transportation, reading pleasure, emotional involvement and others (cf., e.g., Bálint et al., 2017; Kuijpers et al., 2021). While this already highlights the importance of suspense, suspense turns out to be a notoriously difficult object of study because it can be defined in a range of categorically different but obviously related ways. For example, suspense can be viewed as a characteristic of the reader's experience or as a characteristic of the text or the plot. When focusing on the former, a range of emotional, cognitive and physiological components become relevant: '[...] a heightened sense of anticipation, a need to know what comes next, often the fear that the "wrong" thing will happen, and all this sometimes accompanied by increased heart rate and bated breath' (Yanal, 1999, p. 128). Yet, it is not clear that a unique emotional or cognitive profile of *suspense* even exists. By contrast, focusing on suspense as a textual property reveals that lexical or grammatical features as well as discourse-structural pragmatic interpretive features (e.g., Carroll, 1996; Köppe and Onea 2023) interact with suspense. Again, while there may or may not be a unique textual profile one can associate with *suspense*, it is beyond doubt that it is the text that causes the emotional and cognitive phenomena experienced by the reader.

Whatever perspective one may have on the ontological nature, triggering and adequate analysis of narrative suspense, an empirical study will invariably require to ascertain the empirical level of suspense as an index of the reading experience. While this measurement itself may be dependent on how exactly suspense is defined, e.g. in terms of an emotional or more of a textual or cognitive notion, one should expect some level of correlation between the measured results irrespective of which perspective is taken: after all, we do expect a causal relation between the two. The goal of this study is to introduce a novel experimental method to ascertain the suspense arc of a written narrative text with high granularity and explore the status of the elicited suspense arc with respect to the way in which the term *suspense* is understood by participants as well as theoretical expectations based on a narratological analysis of the narrative text.

Previous studies captured various notions of suspense like arousal and fear by measuring biological functions like heart rate and skin temperature (Zillmann et al., 1975) as well as pupil size and blinks (Kaakinen & Simola, 2020), pulse volume amplitudes and skin conductance levels (Bente et al., 2022), or using fMRI (Lehne et al., 2015). However, these biological signals also express similarly emotionally aroused states like alertness or curiosity and can only be distinguished with conscious reasoning. Therefore, these kinds of biological measurements do not suffice to identify and measure the development of suspense in a narrative alone. Hence, a specific measure of the conscious self-assessment of suspense by participants will be necessary.

This presents a conundrum: If study participants read the whole text and rate it for suspense in the end, no fine-grained knowledge about the development of suspense within the text can be obtained (see also Hurlburt and Schwitzgebel, 2011, pp. 14–17). If, however, suspense levels are rated after every sentence, the reading experience might be too disrupted for suspense to arise. Thus, readers face the challenge of simultaneously juggling two cognitive tasks: comprehending the text and providing information about the experienced suspense.

Many researchers have tried to overcome this problem in a number of ways. The existing methods for measuring suspense – and more broadly, tension in the subjective experience of narratives – fall into two types: static, carried out after the stimulus was presented, and dynamic, measured while the stimulus is being presented (see below). The static measurement usually relies on a post-event judgment which can in principle pertain both to suspense understood as a cognitive-interpretive phenomenon and to suspense understood as an emotional-cognitive phenomenon. By contrast, the dynamic measurement by its very nature usually pertains to physiological or emotional components that can be measured during the act of reading or exposure to the narrative and usually does not involve any judgment by the participant.

Dynamic measurements are usually carried out in a continuous manner, i.e., by constantly monitoring some features that can be associated with suspense. Obviously, continuous measurements have nearly infinitely high granularity, only constrained by the measurement device resolution. Continuous measurements have been taken for biological signs of tension, suspense and fear e.g. in music (Farbood, 2012; Krumhansl, 1996; Lehne et al., 2014; Madsen & Fredrickson, 1993; Vines et al., 2006), film suspense (Bezdek, 2012; Bezdek & Gerrig, 2017; Bezdek et al., 2015), emotional affect (Ruef & Levenson, 2007), and audio renditions of literary stories (Kaakinen & Simola, 2020; Riese et al., 2014). However, whenever the conscious experience of suspense was asked for, it was again not on a continuous level. For example, Riese et al. (2014) combined their continuous measurement of pupil size with discrete ratings of suspense by the same participants. This way, an indirectly continuous measurement could be achieved that exploited the (not very strong) correlation between the static and continuous measurements. Similarly, Lehne et al. (2015) collected discrete suspense ratings of their target text divided into around 60 segments while recording (continuous) fMRI data.

Since we are interested in a direct measurement of suspense as judged by participants, we will by the nature of our goal be constrained to static measurements. However, we will suggest a method that at the same time reaches a level of granularity comparable to continuous measurement approaches and overcomes the problem explicated above (that increasing the granularity of measurement in a static setting makes the task increasingly disturbing and thus makes the experience of suspense hard if not impossible). We therefore will refer to our paradigm (to be introduced below) as a *hybrid* measurement paradigm.

One major issue for static measuring methods concerns the nature of the elicited data. One scale of measurement includes relative measures, where each new segment is ranked as "more/equally/less suspenseful" than the previous one (Doust, 2015). The challenge of this kind of measurement is that even though the points of interest are chosen consecutively through the story, uneven rises and falls in suspense are impossible to capture since the rating only allows the direction of up or down based on the previous part, but due to limited points of the scale, the intensity of how much it goes up or down is, by definition, always equal. Such methods would work best with a very high granularity measure, as sustained increase and decrease could be indicative of the intensity of decrease or increase, but increasing the granularity usually runs into the problem raised above that the experimental tasks becomes less adequate for judging suspense because it would disrupt the reading experience even more.

A plethora of studies employ absolute measures like Likert scales with varying degrees of granularity: 5-point (Cheong & Young, 2014), 6-point (Knobloch et al., 2004), 7-point (Gerrig & Bernardo, 1994; Hoeken & van Vliet, 2000), as well as the 9-point rating pictographic scale grounded on the SAM model (Delatorre et al., 2018). Some resort to arbitrary fixed numbers, such as 1–100 (Comisky & Bryant, 1982), 1–10 (Lehne et al., 2015), or 1–11 (Riese et al., 2014). These fixed scale measurements face inherent challenges, most notably the ceiling effect. Specifically, it becomes logistically unfeasible for participants to provide accurate ratings since the thorough evaluation of a narrative's every point is prerequisite to identifying the most suspenseful moment. In an environment that is static yet not entirely *post hoc*, participants are indirectly compelled to speculate whether the moment they are currently experiencing could be the narrative's zenith, as opposed to rating the suspense of the specific segment. Soliciting an overall suspense rating only at the conclusion of the narrative provides either no or only temporally lagging insight into the suspense progression during the story's unfolding, consequently compromising the granularity or reliability of the data.

Guided by the work of Doust and Piwek (2017), we opted to employ magnitude estimation as our measure of suspense. With magnitude estimation, participants are asked to attribute a numerical value to a given stimulus, signifying its intensity. Notably, this method imposes no constraints on the number that can be assigned (Teghtsoonian, 2015). The data harvested in this manner can aptly capture the dynamic nature of suspense as the narrative progresses, rendering the measurement remarkably adaptable and responsive to fine-grained changes in suspense intensity (see also Doust, 2015, p. 193). In addition, it eliminates the risk of participants maxing out their measurement scale before realizing that the narrative continues to grow more suspenseful.

Previous studies have evaluated suspense across a spectrum of narratives, ranging from automated system-generated stories (Cheong & Young, 2014; Doust & Piwek, 2017), student-written short stories (Knobloch et al., 2004), and brief narratives (Delatorre et al., 2018;

Zwaan et al., 1995), to the inaugural chapter of a suspense novel (Hoeken & van Vliet, 2000), a shortened and slightly lexically altered suspenseful story (Lehne et al., 2015) and scenes excerpted from suspenseful novels (Bálint et al., 2017; Gerrig & Bernardo, 1994; Riese et al., 2014). Yet, these studies relied on stimuli that were either automatically produced, tailor-made for the experiment, or too short and fragmented to be categorized as suspenseful literature. To gain deeper insights into the processes that generate and shape suspense in the context of literary narrative interpretation, it is beneficial to analyze authentic literary works. Such works follow a genuine suspense trajectory, initiating with a subdued level of suspense and possibly culminating with a decline in suspense. Given the evolving nature of suspense and its dependence on immersive experience, we have opted for a complete literary short story. This choice provides a balance between the requirement for a narrative long enough to allow for suspense to escalate, and the feasibility concerns related to experiment duration. Consequently, we can capture the subtleties of suspense progression, as generally recognized by both narratologists and lay readers, within the practical confines of our study design.

To summarize, assessing suspense in texts presents a multifaceted challenge. Excessive requests for participant ratings can shift their focus from the suspense to the task, yet infrequent requests can lead to sparse data and a potential overlook of suspense fluctuations. The length of the text further complicates matters. While long texts cannot feasibly be rated in their entirety, short texts or those constructed for experiments fail to authentically represent suspense in genuine literary contexts. We aim to bridge this research gap with our simple yet precise, adaptable, and fine-grained measurement method which enables participants to focus on their reading experience with minimal disruption, while intuitively rating suspense throughout the reading process. Thus, we can gather valuable behavioural data in the process of examining authentic, literary short stories. Simultaneously, we aspire to scrutinize the reliability of our measurement by exploring the correlation, if any, between suspense as an emotional-cognitive experience and suspense as an interpretive feature of the text in our experimental context.

EMPIRICAL STUDY

GOALS

The goal of the study is to establish a new experimental paradigm to determine the suspense arc of a written literary narrative with minimal disruption of the reading process, with high granularity and based on the introspective judgment of participants. In order to achieve this, we use a combination of magnitude estimation measurement, an innovative technical design of digital stimuli presentation and questionnaire including a range of questions about the way in which participants interpreted their task, the term 'suspense' and their reading experience.

METHOD

General information

34 participants (aged $\mu = 27.5$, $\sigma = 8.8$, no gender information collected), speakers of German, have participated in a lab study at the University of Göttingen. The participants were acquired via the online platform of the university and a data base for people interested in taking part in experiments. Regarding the formal education level of participants: 20 participants had passed high school exams, 4 held a bachelor's degree, 7 a master's degree, 2 a doctorate and 2 did not give any information. Participants also provided data on their reading habits: on a scale from 0 (=very little) to 4 (=a lot), the mean indicated reading amount was 2.686 (SD = 0.867). None of the participants had previously read the story under investigation. A session took up to 60 minutes. They were remunerated with 12 Euro per hour.

Timeline

The study comprised three stages. In the training and introductory phase, participants were acquainted with the software paradigm, the keyboard and mouse controls of the software and had the opportunity to ask questions. In the main phase, on a computer screen, participants read the target story, "The Brazilian Cat" by Arthur Conan Doyle, translated into German by Rudolph Rocholl (Doyle, 1965), and simultaneously drew a line that represented their felt

suspense, as explained below.[1] The German translation of the text has a total length of 7451 words (Doyle, 1965), which were automatically broken into 447 lines. Due to a small software flaw, the measurement of the last six lines of text was not saved. Starting with line count 0, the analysis comprises 440 lines of text. Finally, in the closing part, participants had to answer a series of questions.

Introductory phase

In the initial phase of the study, participants are requested to fill out a brief questionnaire (see Appendix), capturing information such as their age, level of formal education, handedness and reading habits, including how frequently they engaged in reading and which genres they preferred. In order to account for potential variations in the participants' ability to complete tasks within average times or perceive color changes in our study design, we also collect information about possible dyslexia and dyschromatopsia. We further enquire whether the participants have previously encountered the story to be evaluated.

In the subsequent training phase, participants become familiar with the tasks of manipulating a line on the screen within the given software environment, such as zooming in and out, navigating back and forth, and making corrections. This phase goes beyond written instructions, including interactive, game-like performance tests. These tests are designed to confirm that participants have sufficiently grasped the controls and instructions before progressing to the main task. The introduction also outlines the experimental task: participants are asked to assess the suspense they experience throughout their reading process.

Suspense line elicitation

The visual paradigm includes a coordinate system with the number 0 in the lower left corner and the number 10 in the upper left corner. The participants' task is to draw their felt intensity of suspense with the mouse in this coordinate system while keeping in mind that the value of each point on the line is to be interpreted in the interval between the bottom and maximal number of the current screen.

Below the coordinate system, 10 lines of the text are visible in a text-box. While the suspense line is drawn in the coordinate system, the text until the word that corresponds to the current end of the line changes color in the text box. This way, participants know exactly which part of text on the screen they are currently rating.

At the bottom of the screen, there is a progress-bar like box showing a summary of the line drawn so far and the maximum value used for rating so far – for screens yet to be seen, there is no line visible. This way, participants can see their previously drawn line as well as get an overview of how far they are in the text. The visual environment is shown in Figures 1 and 2.

While reading, participants draw a suspense line. The drawing task is minimally forced in that participants cannot go to the next screen without having given suspense ratings for the whole text section of the screen. This ensures a rating for every part of the story. The cursor can only stay inside the coordinate system. Moreover, the cursor is constrained to forward movement, thus ensuring that at any point participants cannot have more than one pixel corresponding to one position in the text. To navigate the text, participants press the key S to go forward, which automatically saves the currently drawn suspense line data on word level. On each screen (except the very first), the last few millimeters of the line from the previous screen is shown in order to enhance continuity. To adjust the current piece of suspense line, participants can use the key D for zooming out (the number on the scale in the upper left corner will be doubled) and E for zooming in (the number on the scale in the upper left corner will be halved). Zooming has the function of avoiding ceiling effects, as the participants can always zoom out in order to access higher suspense values than available on the screen. During zooming, the absolute size of the drawing box remains constant and thus the change of scale is only reflected in a flattening or steepening of the current shape of the suspense line. By pressing key A, participants are able to go back and consult the text segments they have already rated,

---

1   The choice of the text was grounded in its narratological analysis as shown below in the general discussion section. In particular, we wanted to have a longer text that exhibits non-linear suspensefulness, i.e., a text for which, based on standard narratological theories of suspense, one would expect several suspense peaks throughout the text, as this creates a much better environment for establishing the reliability and validity of the method.

which ensures a natural reading experience as well as the opportunity for them to back-check previously rated suspense levels. However, in order to prevent participants from second-guessing their own rating and then going back to the beginning to start again (which would overwrite their first, intuitive impression as well as extend the time uncontrollably), they are not allowed to *change* their previously recorded measures. The only part of the line they can erase is the one of the actual text segment, in case they misdrew it unintentionally. At the very first screen, the progress bar is not yet visible, as shown in Figure 1. But after finishing the first screen, the progress bar appears, as can be seen in Figure 2. Participants have no direct control over the progress bar, which is automatically updated after each screen. By pressing the key Q, participants can abort the study at all times.[2]
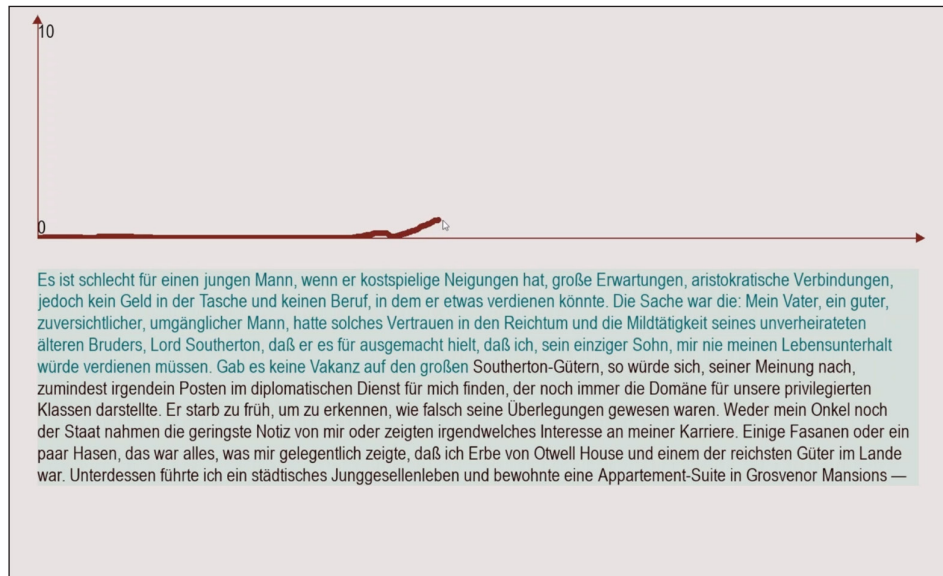
Figure 1 User interface while drawing suspense line at the very beginning. The turquoise portion of the text is already rated for suspense while the black text is not.

Figure 2 User interface while drawing suspense line. (Note the progress line below the text box and the changed coordinates at the upper left corner of the line drawing box).

Questionnaire

After participants finish drawing the line, they are asked to fill in a questionnaire. The questions are answerable with a slider on a scale from –10 (disagree) to 10 (agree). Their abbreviation used in the statistical analysis below as well as their English translations are listed in Table 1. At the very end, participants were given open text fields to report on potential problems with the study setting and to give any other remarks or ideas about their understanding of the German word *Spannung* ('suspense'). They were also given the opportunity to give examples of suspenseful stories that come to their mind. These questions are listed in the Appendix.

---

2   A short video to demonstrate the procedure of data collection can be found under: https://osf.io/m524v/?view_only=a2ef57e599b9483d89ae89e858180a08. The software can easily be adjusted to be used with e.g. a joystick or, in order to be intuitively used by senior or junior participants, a touchscreen. In order to transfer the method to a tablet, some minor re-coding would be necessary.

ABBREVIATION	ENGLISH TRANSLATION
story_int	I found the story interesting
easy_draw_line	It was easy to depict the perceived suspense through the line
distract_line	I was distracted by having to draw a line
diff_match_txt_line	It was difficult to align the line with the text
no_susp_draw_line_as_should	Although I didn't find the text suspenseful, I could see where it should be and drew the line accordingly
no_susp_draw_line_as_expct	Although I didn't find the text suspenseful, I drew the line as I thought the text would expect from me
read_fun	Reading the story was fun
line_as_authr_intnd	I drew the line as the author might have intended the suspense
judg_susp_txt_strctr	I judged the suspense based on the text structure
spont_feel	I judged the suspense based on my spontaneous feeling
line_avg_prsn_draw	I drew the line as I think an average reader would have
line_exp_readr_draw	I drew the line as I think an experienced reader (expert) would have
diff_concentrate	It was difficult to concentrate on reading
difficulties_imagining_scenes	I had difficulties imagining the story
self_Assess_difficult	I had difficulties to evaluate my feeling of suspense

Table 1 Mapping of abbreviations to English translations.

Technical implementation

Our study software was developed using Python and leverages the pygame library (Shinners, 2011). During the suspense line drawing phase, each screen displays ten lines of text. While these lines do not necessarily align with the line length in the printed version of the text, we have maintained the original paragraph breaks. Throughout the process, the height of the drawn suspense line is recorded for each pixel and mapped to the corresponding word on the screen. Consequently, a single word may be associated with multiple suspense-height values even as rated by the same participant, but since they would usually be very close to each other, only the mean value assigned to each word by each participant is used for analysis. Relevant data, including the suspense line and reading time for each screen, are archived in a .csv format. The responses to the initial and concluding questionnaires are also compiled within this .csv file. For statistical analyses, we utilized the R programming language (R Core Team, 2022), all graphs were created using the `ggplot2` package (Wickham, 2016).

RESULTS

In this section, we present the raw results of the study.[3] We provide a more detailed data analysis in the next section as part of our overall discussion of the results including a detailed analysis of the narrative and the role of the answers of the questionnaire with regard to the elicited suspense lines.

In Figure 3, we show the suspense ratings as elicited from the participants for the text at word-level granularity with one standard deviation ribbons on the left-hand side. The extreme variation in the numeric scale used by individual participants can be in principle either due to varied levels of suspense intensity or of the particular way in which participants used the zooming option. Since our focus of interest is the shape of the suspense curve and not so much its individually felt intensity, we applied a z-transformation to the data thus resulting in a much more stable suspense arc for the narrative, as shown on the right-hand side of the figure. At the same time, we reduced the granularity to line level, as 440 data points per participant appear more than sufficient for statistical analysis. All further results are given at line level granularity.

---

3   The data provided by the software is available on "Open Science Framework" using the DOI https://doi. org/10.17605/OSF.IO/YJ2Z7.
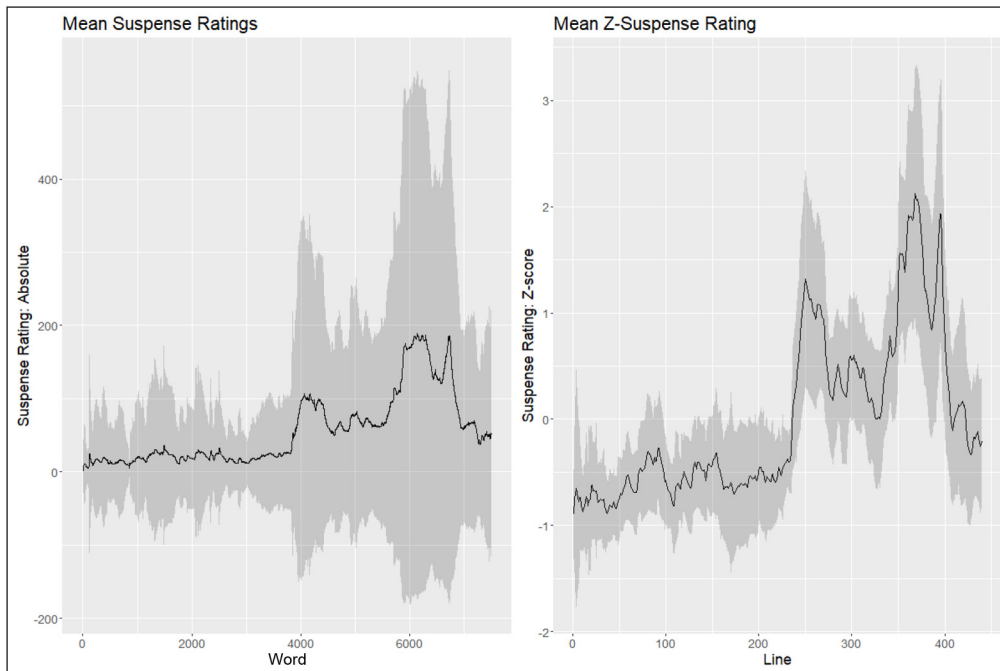
Figure 3 Suspense lines with different levels of granularity.

To investigate the variability in suspense ratings, we applied a linear mixed-effects model. The model included a fixed effect for word number and a random intercept for participant. It was fitted using the `lme4` package in R (Bates et al., 2015). The model summary indicated that the fixed effect of word number was statistically highly significant ($p < 0.001$), suggesting that the elicited data were dependent on the text progression. The proportion of variance in the suspense ratings explained by the model (i.e., the marginal R^2) was obtained using the `r.squaredGLMM` function from the `MuMIn` package (Bartoń, 2023), and was found to be 0.04. Of course, this small value is entirely expected given that no theory of suspense actually predicts a purely linear correlation between text progression and suspense and our text was specifically chosen based on narratological features predicting suspense to not depend only on text progression. To examine the consistency of the suspense ratings across different lines depending on participants, we calculated the Intraclass Correlation Coefficient (ICC) using the `performance` package in R (Lüdecke et al., 2021). The adjusted ICC was estimated to be 0.50 and the unadjusted ICC was 0.48. This suggests a high variability of the suspense lines between participants. We then computed a correlation matrix for the lines elicited by the participants. The average correlation was 0.58 ($\sigma = 0.30$, CI = 0.55–0.60),[4] which can be seen as a reasonable correlation between the individual lines.[5] In assessing the reliability of participants' suspense lines with the omega function from the `psych` package (Revelle, 2023), Omega Total (McDonald, 1999; Revelle & Zinbarg, 2009) was found to be exceptionally high at $\omega_t = 0.99$, indicating a robust consistency across all suspense ratings. Additionally, the Omega Hierarchical value was 0.81, suggesting that a significant proportion of the variance in the ratings is due to a shared underlying perception of suspense. Complementing these findings, Cronbach's Alpha confirmed the internal consistency with a value of 0.98 (Cronbach, 1951).

As for the questionnaire data, we considered the questions with scalar answers summarized in Table 1. We do not report on the open questions concerning the study setting as they are orthogonal to the aim of this paper. A qualitative analysis of the definitions of suspense is provided in the Discussion. We performed a cluster analysis using the R-package `cluster` (Maechler et al., 2022) with 2 clusters of participants solely based on the answers participants

---

4   To determine the 95% confidence interval (CI) for the average correlation, we employed a bootstrapping approach with 1.000 iterations of resampling with replacement from the observed set of correlations.

5   One reviewer pointed out a potential misunderstanding that is worth noting: while correlations are known to only stabilize with a higher sample size, in our case it is not the sample size but the data points in the individual suspense lines that are used to create the correlation values in the correlation matrix. Hence, given that for every participant we used more than 400 data points, we expect our correlations in the matrix to be highly stable according to the estimates of Schönbrodt and Perugini (2013).

gave to the questions.[6] In particular, we computed a dissimilarity matrix using Gower's distance (Gower, 1971), and used it to perform hierarchical clustering using the complete linkage method. As a result, we ended up with two clusters of participants. In order to be able to interpret the clusters, we provide a detailed breakdown of the answers given by the participants in each of the clusters in Figure 4, ordered according to their discriminatory power between the clusters, i.e., the questions on the left more significantly discriminate the clusters from each other. Thereby the solid sign represents the grand mean for each question across the two clusters.
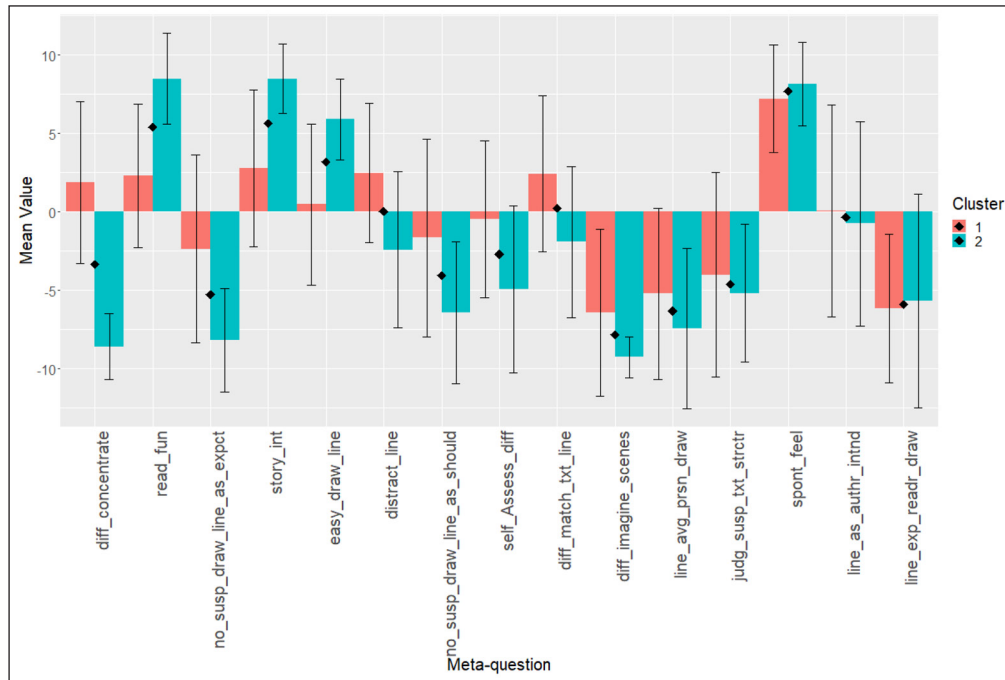
Figure 4 The answer means and standard deviations given to the questions for each cluster of participants, ordered by discriminatory power (strongest to weakest).

Based on this, it is easy to intuitively characterize the two groups of participants as follows:

1. Distracted by the task, less interested in the story and encountered more difficulties imagining the scenes.

2. Engaged and interested, not distracted by the task and had fun during the study.

With this in mind, it appears that cluster 2 comprising 14 of the participants can be roughly understood as persons whose judgment should be closest to what the textual suspense line we wish to capture should be. By contrast, cluster 1, comprising 21 participants, arguably had difficulties experiencing suspense when reading the story. There was no significant effect of age of participants regarding the clusters (t = 1.5241, df = 32.106, p = 0.1373) but the mean age of cluster 1 (μ = 28.9) was somewhat higher than the mean age of cluster 2 (μ = 24.8).

We have used the answers to the questions about the education as well as the amount and preference of reading of participants as predictors in several multinomial models using the `nnet` package (Venables & Ripley, 2002) that predict the cluster for each participant based on these factors and found that these variables are highly significant both alone and in interaction: We found that reading more and having a Master or PhD degree (but also not being an undergraduate student in general) were positively correlated with cluster 2. Similarly, reported reading preferences grouped into binary categories had significant effects on participant clusters. For instance, preference for classical literature greatly increased the likelihood of

---

6   The Gap Statistic (Tibshirani et al., 2001) was employed to determine the optimal number of clusters. The analysis was conducted using the `clusGap` function from the `cluster` package (Maechler et al., 2022), with a maximum number of clusters (K.max) set to 10 and 100 simulated reference sets (B) generated. The method 'firstSEmax' with a SE.factor of 1 was used to select the number of clusters. The results of the Gap Statistic analysis indicated that the optimal number of clusters is 2, as the gap statistic values reached a peak at K = 2 (gap statistic = 0.170), and subsequently decreased with further increasing cluster numbers. The gap statistic values ranged from 0.149 to 0.132 for K = 1 to 10, respectively, indicating the potential presence of two natural clusters in the data.

being in cluster 2, and so did, to a lesser degree, preference for fantasy and science fiction. However, we limit the interpretation of these effects given our study's scope and potential for misinterpretation of results due to societal bias.

A qualitative analysis of how participants defined suspense as well as their examples of suspenseful texts, however, showed no clear difference between the clusters, which mirrors the complexity of the phenomenon (as discussed in the Introduction): Both clusters include definitions of suspense as a subjective, mental and/or physical feeling that can be measured objectively; the feeling of being "gripped" or "enthralled" by the text; wanting to keep reading, to read faster and the feeling of even "not being able to read fast enough". It was furthermore described as an immersion into the story and a rooting for characters. Concerning the narrative level, suspense was described as a feeling of unpredictability or the presence of unexpected, surprising events; as expectations towards as well as the need to know about future events of the story and as uncertainty about the resolution of situations. Participants of both clusters also mentioned the presence of danger in the story. Only in cluster 1, the presence of unresolved situations and the need to know the reason for certain events of the story (i.e. a mystery story pattern) was mentioned. Concerning suspenseful stories recommended by the participants, the genres included thriller, crime, fantasy, science fiction and some literary classics.

Summing up, the qualitative analysis suggests that there is no correlation between participants' competence in defining suspense and the degree to which they experience a text as suspenseful. Also, readers may be good at detecting suspenseful passages without experiencing feelings of suspense (see Henning, 2009, p. 213). This is also reflected in the high correlation of the suspense lines of the different clusters shown in Figure 5. The suspense arcs are computed using Locally Weighted Scatterplot Smoothing (LOWESS; Cleveland, 1979) with span 0.01. In particular, we see that the shapes, in terms of z-transformed suspense lines, are very similar among the groups, but there are differences regarding the height of the lines when using absolute values, which we will discuss in more detail in the next section.
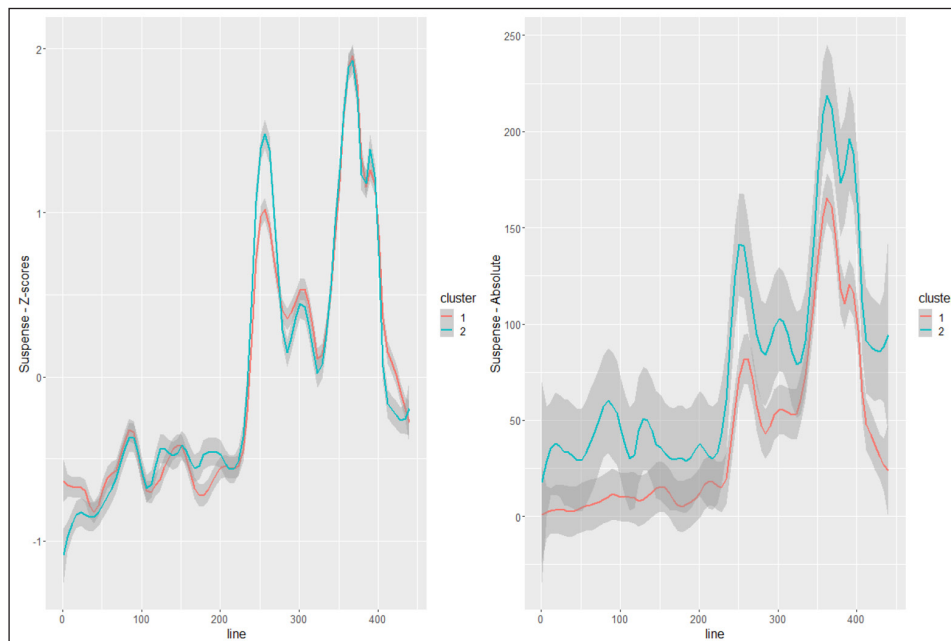
Figure 5 Suspense arc for the two participants' clusters.

DISCUSSION

Our method allows readers to continuously and intuitively assess their personally experienced suspense as they read (online-measure), thus employing introspective self-assessment. This strategy minimizes the cognitive load, potentially facilitating a deeper immersion into the text – thereby yielding more precise representations of the phenomenon under investigation. By employing a line drawing method, as opposed to direct numerical assignment, we could lessen the interruption of the reading process and the intrusiveness of the task to a degree. The act of drawing a line is more intuitive and less distracting than numerical contemplation.

Our observations revealed a high correlation between the suspense lines drawn by different participants. This finding suggests a common understanding of the task amongst participants and a high reliability of our method. Additionally, even when participants were grouped based

on their responses to what we have termed 'meta questions,' the suspense arc for each group retained significant similarity. The average correlation for the second cluster increased to 0.61, while in the first cluster it remained at 0.55. Moreover, the fact that there were no significant age differences between the clusters is indicative of the fact that the method could potentially be used for participant groups of all ages as long as basic familiarity with the hardware environment is given.

Another important question is about the representativeness of the sample of participants. Given the demographic information we collected and data on formal education and reading habits, it should be safe to conclude that our sample is not representative of the overall population, however, it seems possible to generalize our findings to a population with high formal education level and a generally high level of reading experience.

GENERAL DISCUSSION

Even if our sample size of participants is limited and in this paper we have only evaluated the results for a single text, the results show such a high correlation between the suspense lines drawn by the participants when it comes to the shape (and not necessarily the intensity) of the suspense lines that we see sufficient initial evidence for a very high reliability of our method. The main remaining question relates to the validity of our measurement. Construct validity seems to be given since we specifically asked participants to draw suspense lines and since any theory of narrative suspense should be at least very closely related to what we generally mean by the term *suspense* in every-day life. In what follows we provide a narratological analysis of the text and discuss the predictions of narratological analyses and theories regarding suspense in the story. We take narratological theories of suspense to be best suited for establishing *ex ante* which parts of a text should and should not be suspenseful. This we take to be – for better or worse – the gold standard of any suspense measurement, and hence we show how well our measurement aligns with the predictions of such theories. This we take to test the criterion validity of our measurement. Finally, in this section, we also discuss differences between the two clusters of participants established above when it comes to suspense rating, thus gaining evidence about the robustness of our suspense measurement.

NARRATOLOGICAL ANALYSIS

The story belongs to the genre of horror (González-Moreno, 2009; Jones, 2016). It features retrospective narration by a first-person (autodiegetic) narrator (Marshall). Table 2 represents both composition and plot of the story.[7]

Table 2 Story composition and plot.

LINE NUMBER	TEXT PART	PLOT
0–24	exposition	Marshall King is the prospective sole heir of a very wealthy uncle, but currently he is facing bankruptcy.
25–395	main part	Marshall visits the estate of his cousin Everard. Everard treats him favorably while Everard's wife is hostile towards him. Everard promises to help Marshall avert bankruptcy. In a stormy night, both men visit a dungeon-like room on the estate that houses the cage of a large cat of prey. Everard locks Marshall up in the dungeon. Marshall begs for help.
(244)		Everard opens to the cage of the cat of prey, using a mechanism from outside the dungeon. Marshall vainly tries to keep the cage door shut. But the cat of prey does not attack. Marshall manages to climb on top of the cage. The cat attacks and hurts Marshall. He eventually manages to get into the cage and close the door behind him, but the cat wounds him. Once in the cage, he passes out.
(382)		The next morning, Everard enters the dungeon. He is killed by the cat. Servants find Marshall lying unconsciously in the cage and kill the cat.
396–440	ending	Several weeks later, Marshall recovers. He subsequently learns that Everard's wife tried to save him by acting hostilely. Marshall also learns that his wealthy uncle died the very night he spent in the dungeon. It was Everard's plan to kill Marshall in order to inherit the money instead of Marshall.

7   A full plot annotation based on four annotators has been prepared and is available from the corresponding author upon request.

Both the exposition and the ending of the story feature summarizing narration while the main part features isochronous/scenic narration (Rimmon-Kennan, 2002, pp. 51–56). The ending features a plot twist when Marshall learns that Everard's wife actually tried to help him. Since the first-person narrator does not provide this information in the main part of the story, readers are misled concerning the wife's intentions; the story thus qualifies as unreliably narrated in the sense of Stühring (2011).

SUSPENSE THEORY PREDICTIONS ABOUT THE STORY

Current theoretical approaches disagree as to how narrative suspense should be defined. However, many approaches converge in assuming that suspense is a prospective cognitive-emotional state that involves the anticipation of future events in the story that are assigned specific conditional probabilities as well as desirability ratings. In particular, readers deem a passage of text suspenseful if they think that some undesirable event is likely to happen, while they still hope that it does not happen (Carroll, 1996; Carroll, 1990; Henning, 2009, 137f.). This is sometimes dubbed the 'standard account' of suspense, summarized as a combination of hope, fear, and uncertainty (Ortony et al., 1988; Smuts, 2008, 281f.). According to this account, and a bit more formally put, a passage of text is experienced as suspenseful for as long as (a) a future-directed binary question is open ('Will X happen?'), which (b) concerns future events with opposed probability and desirability ratings (e.g., 'It is likely that the protagonist is harmed and unlikely that he is not harmed').

Thus far, the standard account makes a prediction as to whether or not a whole passage of text (not necessarily the entire text) is suspenseful. The text we investigated fits this prediction as follows. The main part of the story brings up the question whether Marshall will be hurt by the cat of prey (line 234). Being attacked by a cat of prey is a natural object of fear (a prospective state that is undesired). The question as to whether the undesired outcome ensues is kept open until line 395, when readers learn that Marshall is eventually safe from further attacks. (Note that this is not yet the end of the story. Thus, the prediction for suspense only holds until the end of this passage.)

Carroll assumes that, in a suspenseful passage of text, it is necessary that "the countervailing probabilities be posed saliently" (Carroll, 1990, p. 141), i.e., we should expect scenes which ostensibly make it likely that the undesired outcome ensues. Carroll also suggests a "formula" according to which probability ratings and degrees of suspensefulness co-vary in that things become more suspenseful if something happens in the story that diminishes the likelihood of the desired outcome (and thus respectively heightens the likelihood of the apprehended outcome). In his example (taken from the film "Night of the Living Dead"), the protagonists try to escape a bunch of zombies by driving off with a truck, and the likelihood of their success is subsequently diminished by adding "probability factors" such as an increase in the sheer number of zombies, difficulties in refueling the truck, and an accidental fire of the truck (ibid.). Thus, the standard account could be said to assume a (*pro tanto*) positive co-variation of fluctuating probability ratings and degrees of suspensefulness. As pointed out above, in "The Brazilian Cat", not only do we find a prolonged passage which brings up the question whether Marshall will be hurt by the cat of prey (lines 234 to 395), thus suggesting that this passage is suspenseful as a whole, but the intermediate story also gives us events which allow for a reading according to which outcome probabilities are shifting such that we should expect an increase [+] or decrease [−] in suspensefulness (see Table 3).

LINE NUMBER	EVENT	CONDITIONAL PROBABILITY (LIKELIHOOD) OF UNDESIRED OUTCOME/SUSPENSEFULNESS
233–253	Marshall wants out but Everard opens cage from outside	+
281–286	Marshall realizes protective candlelight will soon fade	+
287–295	Marshall reaches a relatively safe position	−
299–307	Cat is not aggressive	−
308–310	Protective candlelight fades	+
312–330	Time passes, nothing happens, cat not aggressive	?

Table 3 Probability/ Suspensefulness.

(Contd.)

LINE NUMBER	EVENT	CONDITIONAL PROBABILITY (LIKELIHOOD) OF UNDESIRED OUTCOME/SUSPENSEFULNESS
332–341	Cat becomes aggressive	+
343–347	Marshall tries to close the cage from above	–
349–355	Cat attacks (jumps)	+
356–361	Marshall blinds cat with jacket and keeps closing cage	?
361–363	By mistake, Marshall gets close to the cat	+
364–366	Cat attacks again	+
367–371	Marshall is in cage with door closed behind him	–
380–396	Everard enters the room and is attacked by the cat	?

There are, however, several limitations to the predictive powers of the standard account.

First, events can be read differently as to their effects on conditional outcome probability. For instance, is the fact that Everard walks off (lines 254–255) good or bad, i.e., does it enhance the likelihood of the desired outcome or not? Similarly, it is not clear whether the part in which Everard enters the cage and is attacked by the cat (380–396) has a positive or negative impact on Marshall's fate.

Second, there is no clear prediction as to what happens to the suspensefulness of a passage if no change in outcome probabilities is made salient. This arguably pertains to passages where the protagonist ponders his predicament (lines 312–330). On the basis of the standard account, we could either predict that there is no change in felt suspense (as there is no change in conditional outcome probabilities), or that there is a decrease in suspense (as nothing happens that brings us closer towards the undesired outcome), or that there is an increase of suspense (as still more time runs out for the protagonist). Similarly, it is not clear what predictions should be made when in a very short part of the text positive and negative probabilities fluctuate rapidly such as in lines 356–361, where, on the one hand, Marshall comes dangerously close to the cat while on the other hand at the same time he pursues a rescue plan.

Third, Carroll predicts that in a suspenseful passage of text, changes in outcome probability are made salient. However, there is no independent measure of the salience of changes in outcome probability. All we might be able to say is that, if readers change their probability ratings, then, presumably, something has happened which made them do so.

Finally, it is an open question whether the conditions named by the standard account are indeed sufficient for suspense. For it seems that there are other passages in "The Brazilian Cat" that, arguably, meet the criteria. Marshall's pending bankruptcy may be a case in point here (lines 1–12, see our summary above).

We thus conclude that these deficiencies limit the extent to which our method can be evaluated by way of correlating it with predictions from the standard account. Moreover, the above considerations clearly point to limitations of the explanatory power of the standard account more generally.

Despite the issues with the standard account, we believe that an expert narratological analysis provides the most reliable gold standard for validating a suspense measurement. We thus plotted the suspense line of the text against the predictions of the standard account in Figure 6. Upon visual inspection, the gray region of the text that is predicted to be overall the most suspenseful part of the text is indeed clearly the most suspenseful part. A linear regression analysis was conducted to statistically evaluate the relationship between measured suspense and the designated high suspense region (marked as 'high'). The model, where 'high' is a binary variable indicating whether a line is within the gray suspenseful area, showed a significant effect (high: estimate = 1.268, SE = 0.013, t = 95.89, p < 2e-16). The results yielded an R^2 of 0.374, suggesting that the presence in the high suspense region is a strong predictor of the suspense level in the text. Moreover, in those cases in which there was a clear prediction of increasing suspense, the suspense line does increase very clearly. And for those parts for which a temporary relief is predicted, we visually witness decrease of suspense. To statistically

examine the dynamics of suspense, we analyzed changes in z-suspense line by line for each participant, focusing specifically on sections with predicted increases or decreases in suspense. A linear regression model predicting suspense change by a binary variable 'pred' (1 for predicted increase, –1 for predicted decrease) was fitted. The model revealed a significant impact of the predicted change on actual suspense change (pred: estimate = 0.035, SE = 0.002, t = 15.67, p < 2e-16). However, with an R^2 of merely of 0.016, also the limitations of the standard account to predict the dynamics of suspense in detail appear confirmed. Similarly, for the parts for which the standard account has unclear predictions, it would seem that improvements of the standard account are necessary.
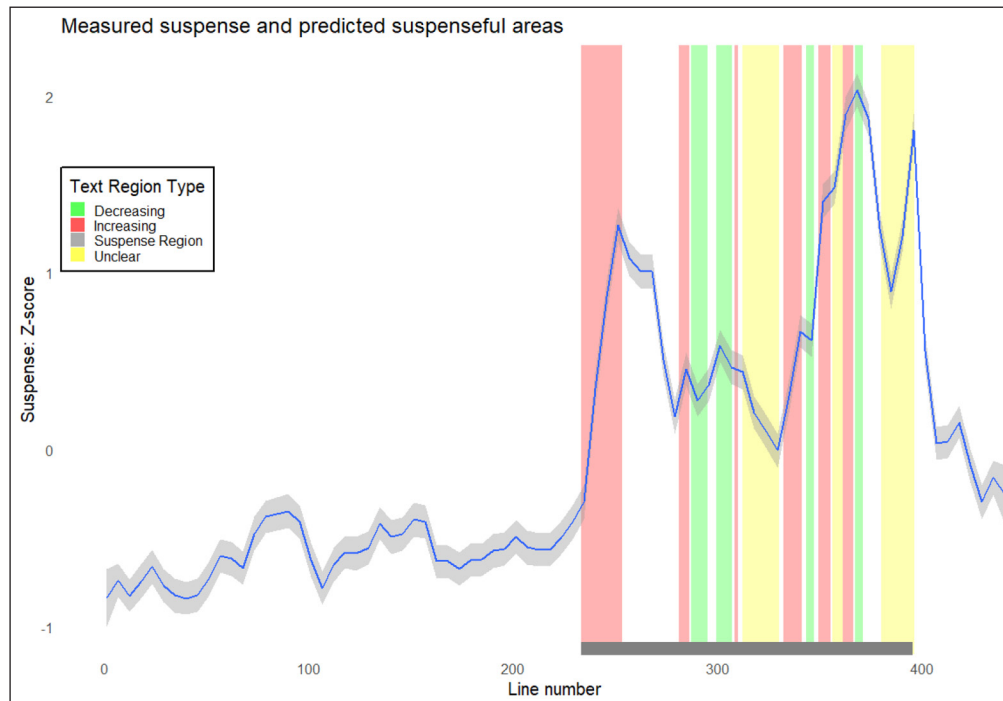
Figure 6 Predicted suspense vs. measured suspense.

HOW ROBUST IS SUSPENSE MEASUREMENT

We conducted a linear regression model to investigate the height difference between the suspense lines across clusters. Our findings indicate that cluster 2 rated suspense higher on average than cluster 1, p < 0.01. Since, however, we are uncertain about the exact meaning of higher intensity given that members of cluster 1 reported to have some difficulty with the line-drawing task, and thus could in principle have used the zooming option less extensively, we can only make the weak claim that some initial evidence suggests that people who found the story more interesting and were not distracted by the task may have indeed experienced suspense to a higher level. Our main focus is, however, on the general shape of the suspense lines and not on their absolute height. Hence, we will rather focus on the z-transformed data. A visual inspection suggests a high alignment between the suspense lines for the two clusters. To statistically verify this, we split the text into chunks of 30 lines, i.e. 3 screens of text, resulting in 15 chunks. We applied a linear mixed-effect model investigating the variability in suspense ratings, including the fixed effect of line and a random intercept for the cluster variable. Unsurprisingly, the fixed effect of line was statistically highly significant in all but the first chunk (p < 0.01). However, the Intraclass Correlation Coefficient (ICC) showed very small values: the highest adjusted ICC was 0.041 and the highest unadjusted ICC was 0.039. This implies that the participants' cluster membership had virtually no impact on their suspense line drawing, as far as the shape of the suspense arc is concerned. These findings highlight the robustness of our method: Despite participants in cluster 1 being distracted by the task or not resonating so well with the story, their suspense lines were highly similar to those drawn by cluster 2 participants, who found the task easy and enjoyed the story. Thus, text-structure oriented suspense measurement can be effective even when participants do not report an emotional effect, suggesting it should be the gold standard in literary suspense studies.

Two of the reviewers of the paper raise an important issue that needs to be addressed at this point. Is the observation that the shape of the suspense arcs for the two clusters barely differ really a sign of robustness of the method or should it rather raise the question what the participants

have actually rated? Specifically, if those participants who did not feel suspense drew the same line as those who did, one may worry that none of the groups actually represented their felt suspense. At closer inspection, however, we consider this worry important but unjustified.

First, we note that the main difference between the two clusters concerned the level of distraction and difficulties during the experimental task. Indeed, members of both clusters reported that they followed their spontaneous feeling during the judgment task. From this perspective it seems to us that using the two clusters to establish the robustness of the method is more than justified, as we deem the convergence of the two groups regarding the shape of the suspense arcs in the face of different levels of difficulty to accommodate the experimental task to be a key aspect of robustness. But, secondly, what seems more problematic is that the two clusters clearly differed with respect to the ratings of the following two statements:

> [S1] Although I didn't find the text suspenseful, I could see where it should be and drew the line accordingly

> [S2] Although I didn't find the text suspenseful, I drew the line as I thought the text would expect from me

We do not consider a positive rating to either of these statements as indicative of a total lack of experienced suspense but rather as a suggesting a reduced suspense level associated with a statement about the sort of strategy used to compensate for this reduced level of felt suspense. Moreover, it is members of cluster 2 who generally gave strongly negative ratings to both assertions, while even for members of cluster 1 the overall mean rating was below 0 for both statements. Hence, there was no group of participants who clearly reported not having experienced suspense at all. Indeed, the observed difference in the overall height of the suspense lines provided by the two clusters of participants would correspond to the gradual difference between the ratings given to S1 and S2. Thus, one could argue that the line elicited does naturally correlate with felt suspense and the two strategies in S1 and S2 are used to compensate for felt suspense yielding (essentially) the same suspense arc. However, for the reasons noted above that participants may have used the zoom function differently in the two clusters, we do not wish to over-interpret the difference in the height of the suspense lines and limit ourselves to the conclusion that our method robustly captures the suspense arc at the text-level, i.e., we concentrate on the shape of the suspense arc only.

At this juncture another, related, worry of two anonymous reviewers of our paper needs to be mentioned, namely that participants may in reality not judge suspense (as an experiential emotional feature) but rather that some might provide a rating corresponding to some naïve and spontaneous narrative analysis based on textual features. The main empirical reason to raise this worry is, again, that the participants in the two clusters appear to have a different experience and yet provide essentially the same line.

While the discussion above addresses part of this worry in suggesting that the key difference between the two clusters appears to be the way in which they dealt with the difficulties imposed by the experimental task and in the intensity of felt suspense rather than its location in the text, the possibility that participants in one or both clusters may have reported on their spontaneous analysis of the text rather than their experience still remains. Indeed, we cannot exclude this possibility entirely and thus leave it to future research to establish the correlation of the measure delivered by our proposed method with more reliable (neuro-)psychological correlates of the emotional components usually associated with suspense. As we have noted above, our method provides the sort of near-continuous measurement needed for such experiments, since psychological correlates alone are hard to interpret lacking fine-grained suspense-measure at the text-level.[8] We do note, however, that this does not contradict our claim that the report is based on self-reflection of participants and mainly on their feelings (as both clusters of participants reported).

---

8   Although our methodology can still be categorized as static — individuals are effectively making *ex post* judgments about the text's interpretation — we argue that its granularity reaches the word level, temporally aligning suspense judgment with suspense experience. Thus, our near-continuous measurement approach can be considered a hybrid between static and dynamic methods, bringing it closer to dynamic measurements. This has the significant advantage that we were able to capture text-based reflexive judgments of suspense, usually ascertained using static models, while having all the advantages of dynamic, continuous, models.

But there seems to be a deeper point here as well: a spontaneous, naïve analysis of textual features may indeed be the empirical ground on which people are generally able to predict suspense and attribute suspensefulness to texts when they do not experience the respective emotional component. Moreover, such a text analysis may be involved even when the respective emotional component is present, as the cognitive side of suspense. We remain neutral with regard to these possibilities, as our data do not give us a way to discriminate between them. But if such a spontaneous text analysis is underlying self-reports of suspense and the experienced emotion of suspense is strongly correlated with reports on such a spontaneous text-analysis, then the connection between these two levels of analysis of suspense is all the more remarkable and would need to be scrutinized in much more detail in future research.

CONCLUSION

Our study presents a novel method to measure suspense in literary narratives that combines the strengths of dynamic and static methods while mitigating their shortcomings. Our measurement approach provides an immersive experience that remains closely aligned with the narrative, bridging the gap between experiential and analytical dimensions of suspense. The method exhibits high reliability as witnessed by a very strong correlation between the suspense lines drawn by the different participants. Moreover, we have mapped the prediction of a narratological analysis of the text, which we consider the gold standard for establishing narrative suspense, against the measurement results and found excellent correlations. Our method's robustness is further exemplified by its performance across diverse clusters of participants. The comparison of suspense lines across groups exhibiting varied levels of distraction and engagement with the narrative and the experimental task demonstrated a high degree of alignment, thus showing that suspensefulness of a text can be measured well using this method even if individual participants may have difficulties with the software or may engage with the text to a lesser degree.

Our approach to suspense measurement stands to open up new avenues of research into suspense and related narratological phenomena. Moreover, this method aligns with contemporary trends towards gamification in scientific research, imbuing the experimental procedure with an interactive, engaging quality. This not only enhances participants' experience but also potentially increases the quality and reliability of the data collected. This study thereby contributes to the burgeoning field of cognitive literary studies, dovetailing with recent advancements in areas such as cognitive narratology (Herman, 2002; Zunshine, 2006) and neuroaesthetics (Chatterjee, 2014).

This Our method makes it possible to acquire an objective measurement of the suspense progression through any given text. The resulting suspense arc, then, can be used to establish correlations with any kind of textual or narrative features like language concreteness (Auracher & Bosch, 2016) or other constructs such as immersion (Jacobs & Lüdtke, 2017; Ryan, 2015), story world absorbtion (Kuijpers et al., 2021), transportation (Gerrig, 1993; Green et al., 2004) and story liking (Jacobs & Kinder, 2019). Future work may explore potential refinements and applications of the method as mentioned above, as well as possibly combining it with neuroimaging techniques like Lehne et al. (2015)'s study to investigate the neural underpinnings of suspense and other narrative emotions.

It is also important to mention some limitations at this juncture, which the reviewers of the paper helped us understand more clearly. In particular, while we took the high correlation between the suspense lines provided by the participants as sufficient initial evidence for the reliability of our method, the study should be replicated and the method applied to other narrative phenomena to further explore its reliability. Similarly, while we have established the validity of our method by mapping the measured suspense to narratological predictions, we readily acknowledge that validity should be further explored in future research both concerning the relation of the suspense measured by the proposed method to other constructs of reading experience and concerning the similarity of our results to those delivered using other methods proposed in prior literature, as discussed in the introduction. Finally, given our limited sample size, our methodological approach regarding clustering of participants based on meta questions could be further explored in studies with larger sample size and thus potentially establish a better theoretical understanding of the factors governing participants' behavior in the task.

To conclude, we have presented a pioneering method for suspense measurement that merges the best aspects of static and dynamic methods, demonstrated its robustness, and proposed its potential for advancing research into narrative suspense. This method, we believe, will play an instrumental role in future narratological studies, forging connections between the intricate structure of narrative texts and the complex experiences they evoke in readers.

APPENDIX

QUESTIONNAIRE BEFORE MEASUREMENT	ENGLISH TRANSLATION
Teilnehmer*innen-ID	Participant ID
Alter	Age
Höchster Abschluss	Highest Level of Education
Wie viel liest du? (z.B. "sehr wenig", "wenig", "durchschnittlich", "viel", "sehr viel")	How much do you read? (e.g., "very little", "little", "average", "a lot", "very much")
Welche Buchgenres liest du?	Which book genres do you read?
Bist du Links- oder Rechtshänder*in?	Are you left or right-handed?
Kennst du "Die brasilianische Katze" von Sir Arthur Conan Doyle?	Are you familiar with "The Brazilian Cat" by Arthur Conan Doyle?
Hast du eine Farbfehlsichtigkeit? (Zum Beispiel Rot-Grün-Schwäche)	Do you have any color vision deficiency? (e.g., red-green color blindness)
Bist du dyslexisch?	Are you dyslexic? (reading and writing difficulties)
QUESTIONNAIRE AFTER MEASUREMENT	ENGLISH TRANSLATION
Ich hatte Schwierigkeiten, mir die Szenerie vorzustellen	I had difficulties imagining the scene
Es war einfach, die empfundene Spannung durch die Linie abzubilden	It was easy to represent the perceived suspense through the line
Ich fand die Geschichte interessant	I found the story interesting
Marshall ist verheiratet	Marshall is married
Es hat mich abgelenkt, eine Linie zeichnen zu müssen	It distracted me to have to draw a line
Es war schwierig, Linie und Text aufeinander abzustimmen	It was difficult to coordinate the line and text
Obwohl ich den Text nicht spannend fand, konnte ich sehen, wo er es sein sollte, und habe die Linie entsprechend gezeichnet	Although I didn't find the text suspenseful, I could see where it should be and drew the line accordingly
Obwohl ich den Text nicht spannend fand, habe die Linie so gezeichnet, wie ich dachte, dass der Text es von mir erwartet	Although I didn't find the text suspenseful, I drew the line the way I thought the text expected me to
Das Lesen der Geschichte hat mir Spaß gemacht	Reading the story was enjoyable
Ich habe die Linie so gezeichnet, wie der Autor die Spannung beabsichtigt haben könnte	I drew the line as the author might have intended the suspense to be
Ich habe die Spannung nach meinem spontanen Empfinden beurteilt	I rated suspense based on my spontaneous feelings
Ich habe die Spannung anhand der Textstruktur beurteilt	I rated suspense based on the text structure
Die Geschichte spielt in Brasilien	The story is set in Brazil
Ich fand es schwierig, mein Spannungsempfinden zu beurteilen	I found it difficult to assess my sense of suspense

QUESTIONNAIRE AFTER MEASUREMENT	ENGLISH TRANSLATION
Es war schwierig, sich auf das Lesen zu konzentrieren	It was difficult to concentrate on reading
Ich habe die Linie danach gezeichnet, wie durchschnittlich erfahrene Leser/innen sie meiner Meinung nach gezeichnet hätten	I drew the line as I thought averagely experienced readers would have drawn it
Ich habe die Linie so gezeichnet, wie erfahrene Leser/innen (Experten) sie meiner Meinung nach gezeichnet hätten	I drew the line as I thought experienced readers (experts) would have drawn it
Hattest du Probleme mit dem Experimentsetting und/oder hast du weitere Anmerkungen oder Ideen?	Did you have problems with the experimental setup and/or do you have any additional comments or ideas?
Wie würdest du "Spannung" definieren?	How would you define "suspense"?
Nenne besonders spannende Geschichten, die du gelesen hast (egal welcher Art)!	Name particularly suspenseful stories you have read (of any kind)!

DATA ACCESSIBILITY STATEMENT

ETHICS AND CONSENT

The study was approved by the Ethics Committee of the University of Göttingen.

ACKNOWLEGEMENTS

FUNDING INFORMATION

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Maria Bentz, Data Curation, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; Maya Cortez Espinoza, Data Curation, Methodology, Software, Writing – Review & Editing; Vesela Simeonova, Data Curation, Methodology, Writing – Review & Editing; Tilmann Köppe, Conceptualization, Funding Acquisition, Methodology, Resources, Supervision, Writing – Review & Editing; Edgar Onea, Conceptualization, Funding Acquisition, Formal Analysis, Methodology, Software, Supervision, Validation, Visualization, Writing – Review & Editing.

AUTHOR AFFILIATIONS

Maria Bentz ⓘ orcid.org/0009-0004-2168-4808
University of Göttingen, DE

Maya Cortez Espinoza ⓘ orcid.org/0009-0008-1776-4986
University of Graz, AT

Vesela Simeonova ![ORCID](orcid.org/0000-0002-3841-1265)
University of Graz, AT

Tilmann Köppe ![ORCID](orcid.org/0000-0001-7923-5432)
University of Göttingen, DE

Edgar Onea ![ORCID](orcid.org/0000-0002-8217-4463)
University of Graz, AT

REFERENCES

Auracher, J., & **Bosch, H.** (2016). Showing with words: The influence of language concreteness on suspense. *Scientific Study of Literature*, 6(2), 208–242. DOI: https://doi.org/10.1075/ssol.6.2.03aur

Bálint, K., Kuijpers, M. M., Doicaru, M. M., Hakemulder, F., & **Tan, E. S.** (2017). The effect of suspense structure on felt suspense and narrative absorption in literature and film. *Narrative Absorption*, 27, 177–197. DOI: https://doi.org/10.1075/lal.27.10bal

Bartoń, K. (2023). *MuMIn: Multi-model inference* [R package version 1.47.5]. https://CRAN.R-project.org/package=MuMIn

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

Bente, G., Kryston, K., Jahn, N. T., & **Schmälzle, R.** (2022). Building blocks of suspense: Subjective and physiological effects of narrative content and film music. *Humanities and Social Sciences Communications*, 9(449), 1–13. DOI: https://doi.org/10.1057/s41599-022-01461-5

Bezdek, M. A. (2012). *Changes in attentional focus during suspenseful film viewing* (Doctoral dissertation). State University of New York at Stony Brook.

Bezdek, M. A., & **Gerrig, R. J.** (2017). When narrative transportation narrows attention: Changes in attentional focus during suspenseful film viewing. *Media Psychology*, 20(1), 60–89. DOI: https://doi.org/10.1080/15213269.2015.1121830

Bezdek, M. A., Gerrig, R. J., Wenzel, W. G., Shin, J., Revill, K. P., & **Schumacher, E. H.** (2015). Neural evidence that suspense narrows attentional focus. *Neuroscience*, 338–345. DOI: https://doi.org/10.1016/j.neuroscience.2015.06.055

Carroll, N. (1996). *Theorizing the moving image*. Cambridge University Press.

Carroll, N. (1990). *The philosophy of horror or paradoxes of the heart*. Routledge. DOI: https://doi.org/10.4324/9780203361894

Chatterjee, A. (2014). *The aesthetic brain: How we evolved to desire beauty and enjoy art*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199811809.001.0001

Cheong, Y.-G., & **Young, R. M.** (2014). Suspenser: A story generation system for suspense. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(1), 39–52. DOI: https://doi.org/10.1109/TCIAIG.2014.2323894

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829–836. DOI: https://doi.org/10.1080/01621459.1979.10481038

Comisky, P., & **Bryant, J.** (1982). Factors involved in generating suspense. *Human Communication Research*, 9(1), 49–58. DOI: https://doi.org/10.1111/j.1468-2958.1982.tb00682.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. DOI: https://doi.org/10.1007/BF02310555

Delatorre, P., León, C., Salguero, A., Palomo-Duarte, M., & **Gervás, P.** (2018). Confronting a paradox: A new perspective of the impact of uncertainty in suspense. *Frontiers in Psychology*, 9(1392). DOI: https://doi.org/10.3389/fpsyg.2018.01392

Doust, R. (2015). *A domain-independent model of suspense in narrative* (Doctoral dissertation). The Open University. DOI: https://doi.org/10.21954/ou.ro.0000ab68

Doust, R., & **Piwek, P.** (2017). A model of suspense for narrative generation. *Proceedings of the 10th International Conference on Natural Language Generation*, 178–187. DOI: https://doi.org/10.18653/v1/W17-3527

Doyle, A. C. (1965). *Gesammelte Werke in Einzelausgaben. Bd. 13: Die brasilianische Katze. Hrsg. von Nino Erné*. Ullstein.

Farbood, M. M. (2012). A parametric, temporal model of musical tension. *Music Perception: An Interdisciplinary Journal*, 29(4), 387–428. DOI: https://doi.org/10.1525/mp.2012.29.4.387

Gerrig, R. J. (1993). *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press. DOI: https://doi.org/10.12987/9780300159240

Gerrig, R. J., & **Bernardo, A. B. I.** (1994). Readers as problem-solvers in the experience of suspense. *Poetics*, 22(6), 459–472. DOI: https://doi.org/10.1016/0304-422X(94)90021-3

González-Moreno, B. (2009). Sir Arthur Conan Doyle and the hauntings of the "American Blood-Curdler". *The Edgar Allan Poe Review*, 10(2), 25–35. DOI: https://doi.org/10.2307/41507877

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*(4), 857–871. DOI: https://doi.org/10.2307/2528823

Green, M. C., Brock, T. C., & Kaufman, G. F. (2004). Understanding media enjoyment: The role of transportation into narrative worlds. *Communication Theory, 14*(4), 311–327. DOI: https://doi.org/10.1111/j.1468-2885.2004.tb00317.x

Henning, T. (2009). *Person sein und Geschichten erzählen: Eine Studie über personale Autonomie und narrative Gründe*. Walter de Gruyter. DOI: https://doi.org/10.1515/9783110210200

Herman, D. (2002). *Story Logic: Problems and Possibilities of Narrative*. University of Nebraska Press.

Hoeken, H., & van Vliet, M. (2000). Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story. *Poetics, 27*(4), 277–286. DOI: https://doi.org/10.1016/S0304-422X(99)00021-2

Hurlburt, R., & Schwitzgebel, E. (2011). *Describing inner experience? Proponent meets skeptic*. MIT Press.

Jacobs, A. M., & Kinder, A. (2019). Computing the affective-aesthetic potential of literary texts. *AI, 1*(1), 11–27. DOI: https://doi.org/10.3390/ai1010002

Jacobs, A. M., & Lüdtke, J. (2017). Immersion into narrative and poetic worlds: A neurocognitive poetics perspective. In F. Hakemulder, M. M. Kuijpers, E. S. Tan, K. Bálint & M. M. Doicaru (Eds.), *Narrative absorption* (pp. 69–96). John Benjamins Publishing Company. DOI: https://doi.org/10.1075/lal.27.05jac

Jones, D. (2016). Introduction. *Doyle, Arthur Conan: Gothic tales*, ix–xxxiii.

Kaakinen, J. K., & Simola, J. (2020). Fluctuation in pupil size and spontaneous blinks reflect story transportation. *Journal of Eye Movement Research, 13*(3). DOI: https://doi.org/10.16910/jemr.13.3.6

Knobloch, S., Patzig, G., Mende, A.-M., & Hastall, M. (2004). Affective news: Effects of discourse structure in narratives on suspense, curiosity, and enjoyment while reading news and novels. *Communication Research, 31*(3), 259–287. DOI: https://doi.org/10.1177/0093650203261517

Köppe, T., & Onea, E. (2023). The nearly missed account of narrative suspense. *Frontiers of Narrative Studies, 9*, 273–289. DOI: https://doi.org/10.1515/fns-2023-2018

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension, and musical ideas. *Music Perception: An Interdisciplinary Journal, 13*(3), 401–432. DOI: https://doi.org/10.2307/40286177

Kuijpers, M. M., Douglas, S., & Bálint, K. (2021). Narrative absorption: An overview. In D. Kuiken & A. M. Jacobs (Eds.), *Handbook of empirical literary studies* (pp. 279–304). De Gruyter. DOI: https://doi.org/10.1515/9783110645958-012

Lehne, M., Engel, P., Rohrmeier, M., Menninghaus, W., Jacobs, A. M., & Koelsch, S. (2015). Reading a suspenseful literary text activates brain areas related to social cognition and predictive inference. *PloS ONE, 10*(5). DOI: https://doi.org/10.1371/journal.pone.0124550

Lehne, M., Rohrmeier, M., & Koelsch, S. (2014). Tension-related activity in the orbitofrontal cortex and amygdala: An fMRI study with music. *Social Cognitive and Affective Neuroscience, 9*(10), 1515–1523. DOI: https://doi.org/10.1093/scan/nst141

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60). DOI: https://doi.org/10.21105/joss.03139

Madsen, C. K., & Fredrickson, W. E. (1993). The experience of musical tension: A replication of Nielsen's research using the continuous response digital interface. *Journal of Music Therapy, 30*(1), 46–63. DOI: https://doi.org/10.1093/jmt/30.1.46

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., Gonzalez, J., Kozlowski, K., Schubert, E., & Murphy, K. (2022). *Cluster: Cluster analysis basics and extensions* [R package version 2.1.4]. https://CRAN.R-project.org/package=cluster

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press. DOI: https://doi.org/10.4324/9781410601087

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511571299

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

Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.3.6]. https://CRAN.R-project.org/package=psych

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145–154. DOI: https://doi.org/10.1007/s11336-008-9102-z

Riese, K., Bayer, M., Lauer, G., & Schacht, A. (2014). In the eye of the recipient: Pupillary responses to suspense in literary classics. *Scientific Study of Literature, 4*(2), 211–232. DOI: https://doi.org/10.1075/ssol.4.2.05rie

Rimmon-Kennan, S. (2002). *Narrative fiction: Contemporary poetics*. Routledge. DOI: https://doi.org/10.4324/9780203426111

Ruef, A. M., & Levenson, R. W. (2007). Continuous measurement of emotion: The affect rating dial. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 286–297). Oxford University Press. DOI: https://doi.org/10.1093/oso/9780195169157.003.0018

Ryan, M. L. (2015). *Narrative as virtual reality 2: Revisiting immersion and interactivity in literature and electronic media*. Johns Hopkins University Press.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. DOI: https://doi.org/10.1016/j.jrp.2013.05.009

Shinners, P. (2011). Pygame. http://pygame.org/

Smuts, A. (2008). The desire-frustration theory of suspense. *The Journal of Aesthetics and Art Criticism*, *66*(3), 281–290. DOI: https://doi.org/10.1111/j.1540-6245.2008.00309.x

Stühring, J. (2011). Unreliability, deception, and fictional facts. *Journal of Literary Theory, 5*(1), 95–10. DOI: https://doi.org/10.1515/jlt.2011.008

Teghtsoonian, M. (2015). Sensation and perception: Direct scaling. *International Encyclopedia of the Social & Behavioral Sciences*, 604–606. DOI: https://doi.org/10.1016/B978-0-08-097086-8.43087-4

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B Statistical Methodology, 63*(2), 411–423. DOI: https://doi.org/10.1111/1467-9868.00293

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer. DOI: https://doi.org/10.1007/978-0-387-21706-2

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition, 101*(1), 80–113. DOI: https://doi.org/10.1016/j.cognition.2005.09.003

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org. DOI: https://doi.org/10.1007/978-3-319-24277-4

Yanal, R. J. (1999). *Paradoxes of emotion and fiction*. Penn State University Press. DOI: https://doi.org/10.1515/9780271071411

Zillmann, D., Hay, T. A., & Bryant, J. (1975). The effect of suspense and its resolution on the appreciation of dramatic presentations. *Journal of Research in Personality, 9*(4), 307–323. DOI: https://doi.org/10.1016/0092-6566(75)90005-7

Zunshine, L. (2006). *Why we read fiction: Theory of mind and the novel*. The Ohio State University Press.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*(5), 292–297. DOI: https://doi.org/10.1111/j.1467-9280.1995.tb00513.x

Scientific Study of Literature is a peer-reviewed open access journal published by International Society for the Empirical Study of Literature.