# Mere plausibility enhances comprehension: The role of plausibility in comprehending an unfamiliar scientific debate

Johanna Abendroth and Tobias Richter

University of Würzburg

**Accepted for publication in the Journal of Educational Psychology (2020)**



Author note

The research reported in this article was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, grant RI 1100/5-3). We would like to thank Andreas Wertgen and Veronika Krawiec for their help in preparing stimulus materials and collecting data. Special thanks go to Jens-Ulrich Maier for his support in preparing the video material. All text materials (translated into English) and the experimental videos (in German) as well as the data files and analysis scripts for the full analyses are available in the Open Science Framework (osf.io/rvg6b) and can additionally be provided by the authors upon request.

Corresponding author:
Johanna Abendroth
University of Würzburg
Department of Psychology, Educational Psychology
Röntgenring 10, 97070 Würzburg, Germany
E-mail: johanna.abendroth@uni-wuerzburg.de

**Abstract**

Readers confronted with unfamiliar and controversial scientific debates tend to rely on simple heuristics such as the perceived plausibility to focus their cognitive resources on specific information during comprehension. In the present experiment, we tested the assumption that plausibility judgments as an integral part of comprehension are used as a simple heuristic to distribute cognitive resources to controversial texts, leading to a better comprehension of information judged as plausible. To experimentally vary perceived plausibility, participants ($N$ = 54 university students) watched one of two video versions on the controversy of spider silk. The videos provided identical factual information but took opposing argumentative claims on the issue (pro vs. con). Afterwards, participants read two conflicting texts (pro vs. con) on the same issue. Plausibility judgments and comprehension for the texts were assessed. In line with the hypothesized mediation model, results revealed that the belief manipulation (i.e., the video versions) affected the perceived plausibility of the controversial texts, which in turn influenced the comprehension of the two texts. The effect of the belief-manipulation, that is, participants' better comprehension of the text that took the same argumentative stance as the video, was fully mediated by perceived plausibility. These results are relevant for educational interventions to improve the comprehension of controversial but unfamiliar scientific studies and for theories on the role of plausibility in (multiple) text comprehension.

*Keywords*: plausibility, validation, beliefs, multiple texts

## Educational Impact and Implications Statement

Readers often use superficially created beliefs to evaluate the plausibility or credibility of statements from controversial texts on unfamiliar topics as part of normal comprehension. The consequence is a better understanding of information that is judged as plausible but a reduced understanding for information judged as implausible. In times of misinformation and fake news spread throughout the World Wide Web, such preferential processing of information perceived as plausible because it is consistent with one's beliefs hampers readers' ability to fully understand and evaluate complex socio-scientific issues, to form well-justified argumentative positions, and to make informed decisions.

## Introduction

The Word Wide Web has revolutionized the way in which individuals seek scientific information and also how scientists present and discuss their academic work. In particular, new theories and empirical results often quickly become available on the World Wide Web. For example, the effectiveness of new medical developments such as artificial nerves made from spider silk is currently presented and controversially discussed in online reputable newspapers (Schwenkenbecher, 2019) or in TV science programs (3Sat, 2018). A similar discussion of a scientific issue can be currently observed on the topic of the COVID-19 pandemic. This transparency is beneficial for scientists because it increases the academic productivity and debate. However, lay people are often overstrained because their prior knowledge is not sufficient to fully understand complex scientific debates. This problem is aggravated by the fact that lay people are usually unaware of their inability to fully understand and evaluate complex scientific debates (Keil, 2010), especially when scientific information is presented in a seemingly easy way (Scharrer, Stadtler, & Bromme, 2014). Consequently, individuals confronted with unfamiliar and controversial scientific information often neglect to evaluate and critically reflect on the information. Instead, they tend to use simple heuristics such as consistency with prior knowledge or beliefs to evaluate the plausibility or credibility of statements (e.g., Britt, Richter, & Rouet, 2014; von der Mühlen, Richter, Schmid, Schmidt, & Berthold, 2016).

The present study examines the assumption that readers use fast and implicit plausibility judgments based on their beliefs to allocate cognitive resources during the comprehension of unfamiliar controversial debates, which in turn affects comprehension of the text information. Plausibility has been defined as the "acceptability or likelihood of a situation or a sentence describing it" (Matsuki et al., 2011, p. 926), "the degree of fit between a given scenario and prior knowledge" (Connell & Keane, 2006, p. 98), or as "what is perceived to be potentially truthful

when evaluating explanations" (Lombardi, Nussbaum, & Sinatra, 2016, p. 35). These definitions all include the concept that plausibility judgements rely strongly on individuals' subjective perception of potential truthfulness. In contrast to objective truth judgements, the information perceived to be plausible during reading might differ among individuals. Hence, a judgment of plausibility can be seen as the assessment of how well a new piece of information coheres with readers' prior knowledge, beliefs, or current understanding of an issue (e.g., Johnson-Laird, 1983; Reder, 1982). Plausibility is therefore not limited to a judgement about the consistency with knowledge. Instead, plausibility perceptions can occur as a consequence of the fit of new information to all information stored in long-term memory and activated during the comprehension of textual information. Activated memory can be background knowledge – when available for a particular topic – but activated memory might additionally be prior beliefs or earlier read information.

In the following, we will first discuss how and when plausibility judgements influence text comprehension, discuss the role of prior beliefs for plausibility judgments in the case of complex and unfamiliar scientific debates, and elaborate on how plausibility judgements affect comprehension outcomes. Afterwards, we report the results from a study that experimentally varied the perceived plausibility of controversial texts about an unfamiliar scientific debate (i.e., the medical use of spider silk) by inducing either pro or contra beliefs in participants. The learning scenario experimentally induced in the experiment strongly resembles typical learning with web-based multiple texts on controversial topics. Plausibility and comprehension were assessed on the level of the individual texts.

**Sensitivity to Plausibility during Comprehension**

Comprehension of a single or of multiple texts include readers' attempt to construct a mental model or situation model of each text, that is, a mental representation of what the text is

about (Kintsch, 1988). The construction of such a representation primarily relies on three passive memory-based processes (O'Brien & Cook, 2016). *Resonance* is the key to activate information from long-term memory. In detail, information encountered during reading a text passively activates information stored in long-term memory (i.e., prior knowledge, prior beliefs, and previous text information) if both types of information overlap sufficiently (e.g., Albrecht & O'Brien, 1993; Beker, Jolles, Lorch, & van den Broek, 2016; Perfetti, Rouet & Britt, 1999). Once information becomes activated, it can be used for *integration*, that is, to connect information from the text with information in memory. Integration is viewed to be mainly based on semantic associations between new information and stored information already known or believed (for an overview, see McNamara & Magliano, 2009). In a third passive comprehension process termed *validation*, readers implicitly evaluate the consistency of new information with activated information from memory (O'Brien & Cook, 2016; Richter, 2015; Richter & Singer, 2017; Singer, 2019). The Resonance-Integration-Validation Model (RI-Val, O'Brien & Cook, 2016) provides an explanation of the parallel asynchronous processing of activation, integration, and validation. In this model, activation occurs first. When sufficient knowledge has been activated during reading, integration starts, which is then followed by validation. Note that the RI-Val Model predicts this triad when reading a particular piece of information before a reader moves on to successive information in the text.

Validation is important during comprehension, especially in the case of conflicting information because it allows readers to "judge whether the information communicated by the various texts is true or plausible" (Richter, 2011, p.126). Strategic judgements about plausibility are inherent in critical thinking and important for conceptual change (for a review on plausibility in conceptual change, see Lombardi et al., 2016). However, as noted in models on text comprehension such as the RI-Val model (O'Brien & Cook, 2016) or the Two-Step Model of

Validation (Richter & Maier, 2017), validation occurs often rather implicitly and without strategic control (for a similar claim on automatically activated plausibility judgments, see Lombardi, Sinatra, & Nussbaum, 2013) -although it can feed into explicit plausibility judgments. Lombardi et al. (2016) argued that plausibility judgements can occur on a continuum ranging from being explicit with a high degree of evaluation to being implicit with a low degree of evaluation. The type of validation described in the RI-Val model (O'Brien & Cook, 2016) or in the Two-Step Model of Validation (Richter & Maier, 2017) would be situated closer to the implicit end of the continuum. Readers judge the plausibility of new information without strategic effort and without much conscious thought, and even if their actual reading task does not require validation. Whether readers rely on such implicit plausibility judgements or continue with more effortful and strategic elaboration that might include more explicit plausibility judgements depends on a variety of person-specific and situational factors such as background knowledge, reading goals, epistemic beliefs, or metacognitive strategies (Richter & Maier, 2017).

Isberner and Richter (2014) provide an overview of reading and reaction time studies as well as eye-tracking studies that investigate how readers notice and react to (im-)plausibility of psycholinguistic information. The authors concluded from the reviewed empirical research that implausible and plausible information is processed differently, even at early processing stages. For example, implausible information often leads to automatic cognitive disruptions such that readers often fixated implausible information longer (e.g., Cook & Myers, 2004), read such information longer (e.g., Albrecht & O' Brien, 1993) and increased their first fixations and gaze durations directly on the critical word (Matsuki et al., 2011). Moreover, readers are sensitive to implausibility such as contradictions within a text or prior knowledge violations as early as they understand the semantic meaning of the stimulus material (e.g., Ferretti, Singer, & Patterson, 2008; Hagoort, Hald, Bastiaansen, & Peterson, 2004; Singer, 2006), even if their task requires no

validation or semantic processing (e.g., Isberner & Richter, 2013; Richter, Schroeder, & Wöhrmann, 2009). Hence, as a core element of validation, implicit and non-strategic plausibility judgements about new information are made during comprehension and are based on information activated from long-term memory. If readers are sensitive to the plausibility of information during initial reading, the outcome of these implicit plausibility judgements can then be used as a heuristic to also focus attention on a more implicit and non-strategic level – a tendency to pay more attention to plausible and less attention to implausible information.

**Prior Beliefs and Plausibility Judgments**

Most research on validation has been conducted with single statements or short stories that provided information inconsistent with earlier parts of the text (e.g. Cook & Myers, 2004; Albrecht & O' Brien, 1993) or violations of general world knowledge (e.g., Isberner & Richter, 2013; Richter et al., 2009). The methods used in these studies were driven by the assumption that plausibility judgments often depend on earlier read information or prior knowledge for a specific topic as epistemic background. However, recent research indicates that readers might also rely on other types of information stored in long-term memory, especially when prior knowledge might not be readily available. For example, prior beliefs are also used by readers to validate new textual information quickly and efficiently (Gilead, Sela, & Maril, 2018; Maier, Britt, & Richter, 2018; Voss, Fincher-Kiefer, Wiley, & Silfies, 1993; Wyer & Radvansky, 1999). For example, Gilead and colleagues (2018) investigated whether readers involuntarily notice inconsistencies of single sentences with their prior beliefs. Participants were instructed to provide judgements about the grammatical accuracy of statements about political, personal, or social issues (e.g., The internet has made people more isolated/sociable). In addition, participants' agreement with the claims was assessed. Participants made faster judgments for grammatically correct statements when they agreed with the claim. In contrast, their grammatical judgment was delayed when the

sentence was inconsistent with participants' beliefs. Hence, similar to the negative response tendency that was found for false or implausible statements (e.g., Isberner & Richter, 2013; Richter et al., 2009), the belief-inconsistency of the grammatically correct statements interfered with participants required "correct" response.

An eye-tracking study by Maier et al. (2018) further supports the idea that prior beliefs are activated initially during comprehension and used for implicit validation. First-pass rereading times for belief-inconsistent information were longer in readers with strong prior beliefs, indicating an early monitoring process that evaluates the belief-consistency of new information. This early processing leads to immediate disruptions and slow-downs in reading (see also Wolfe, Tanner, & Taylor, 2013). Hence, we consider understanding and validating discourse information as a dynamic and interleaved process, during which not only prior knowledge but also prior beliefs might be used as epistemic background to evaluate the plausibility of new information quickly and efficiently.

Using prior beliefs for validation should occur especially for multiple texts with competing claims about scientific issues (Richter & Maier, 2017). The Two-Step Model of Validation (Richter & Maier, 2017) specifies how prior belies affect the comprehension of unfamiliar but belief-relevant multiple texts. Based on the mechanisms outlined above, the Two-Step Model of Validation assumes that two successive steps are influential in the comprehension of belief-relevant multiple texts. Step 1 includes monitoring and detecting belief-consistency as part of routine validation. In the case of belief-relevant information, plausibility and belief-consistency go hand in hand. Readers are assumed to have a tendency to focus their cognitive resources on information judged as plausible during passive validation. If this immediate effect of beliefs on plausibility judgments as a by-product of comprehension is not followed by further strategic attempts to resolve inconsistencies or contradictions, the effect will result in a

preferential processing of plausible or belief-consistent information and a better memory and comprehension for this information (Richter & Maier, 2017). In Step 2, the model proposes that a balanced mental model of a controversial issue can occur when readers engage in strategic and resource-intensive elaboration of implausible or belief-inconsistent information. However, such resource-intensive processing calls for suitable reading goals or motivation that enhances readers' standard of coherence as a crucial point predicting when readers are satisfied with their reading outcome (van den Broek, Beker, & Oudega, 2015). In the next section, we review studies on the effects of plausibility on memory and text comprehension that shed light on the assumptions made by the Two-Step Model of Validation.

**Effects of Plausibility on Memory and Comprehension**

Readers are sensitive to the plausibility of textual information during comprehension. Research suggests that plausibility judgments are used as a simple heuristic to distribute cognitive resources for comprehending information. In other words, the plausibility of information from a text influences the likelihood that this information becomes part of the text's situation model (*plausibility effect*). For example, Schroeder, Richter and Hoever, (2008) found that the plausibility and comprehension of information reciprocally influenced each other, that is that their relationship is bi-directional. Psychology undergraduates read expository texts with plausible and implausible arguments. Implausible arguments were created by inserting argumentation errors in sentences such as contradictions or a conversion of cause and effect. After reading the expository texts, participants were asked to indicate the extent that the paraphrases and inferences as experimental items matched the content of the texts in a verification task (i.e., were part of the texts' situation model) and the extent that the items were convincing in a validation task (i.e., were plausible). The results from an analysis with multinomial models revealed that plausible information was more often perceived as part of the

situation model, that is, the information had a higher likelihood of being verified as belonging to the text's content. In contrast, information that had been integrated into the situation model was more likely judged as plausible regardless of its objective plausibility.

A similar effect of plausibility on the memory for information has been found in simple news stories (de Pereyra, Britt, Braasch & Rouet, 2014) and short everyday stories (Black, Freeman, & Johnson-Laird, 1986). Plausibility has been also found to be relevant for the memory of logical errors and fallacies. Hinze, Slaten, Horton, Jenkins, and Rapp (2014) investigated the effects of plausibility on the use of misinformation to answer factual test questions. The authors found that readers used factual misinformation to answer test questions to a greater extent when the misinformation was plausible than when it was implausible (close to zero commission errors). Think-aloud data from this study further suggest that implausible misinformation led to more strategic skeptical responses and less strategic acceptance responses compared to plausible misinformation. No such effect of strategic processing was found on the use of plausible misinformation, which is viewed by the authors as a lack of active skepticism toward plausible misinformation. Nevertheless, this finding might further indicate that implicit processes (which are not observable by think-alouds) influenced the greater use of plausible compared to implausible misinformation because plausibility perceptions can occur both as a result of implicit processing with less strategic control of the individual (e.g., Richter, 2015) or as a result of more explicit processing with more strategic control (e.g., Lombardi, et al., 2016).

In the context of multiple text comprehension, Maier and Richter (2013) found similar evidence for a link between perceived plausibility of information and readers' situation model. In their study, plausibility of information was not varied, but they assessed the subjective or perceived plausibility of information for each participant. Psychology undergraduates read two controversial texts on a recent controversy from educational science and afterwards provided

binary responses to paraphrase, inferences, and distracters in a recognition task (to assess the memory for each text), a verification task (to assess the situation model strength for each text) and a validation task (to assess plausibility judgements for each test item, i.e., sentences). They found a plausibility effect for inferences. Inferences that were perceived as plausible by readers were more likely integrated into readers' situation model of the text compared to inferences perceived as implausible. Similarly, information perceived as plausible by readers also had a higher chance of becoming part of their memory for text.

In sum, ample evidence has shown that plausibility judgements play a critical role during comprehension and have a strong biasing influence on comprehension outcomes such as memory for text and the texts' situation model. Studies on validation, however, that have experimentally manipulated plausibility have varied the text material between conditions, for example, by inserting logical errors or fallacies (e.g., Schroeder, et al., 2008; Hinze et al., 2014). The effects of such manipulations might depend on whether readers possess knowledge about argumentation. Indeed, attempts to use very similar tasks to assess individual differences in argument evaluation have revealed large individual differences (Münchow, Richter, von der Mühlen & Schmid, 2019). The study by Maier and Richter (2013) acknowledged the subjective nature of plausibly judgements. However, the shortcoming of this study is that plausibility was not experimentally varied.

## Rationale and Overview

We endorse the view that (more or less justified) beliefs are used for parallel and automatic plausibility judgments during comprehension, which have an impact on the memory and comprehension of texts. This should be especially important for complex and unfamiliar scientific debates because readers often lack relevant background knowledge to fully evaluate and scrutinize controversial claims for such debates (Keil, 2010). Prior research supports the

assumption that plausibility judgments strongly influence comprehension (e.g., Maier & Richter, 2013; Schroeder et al., 2008). However, plausibility in this research has been either experimentally varied by, for example, including flawed arguments, or it was not experimentally varied but assessed in the form of subjective judgements. The purpose of the present study was to experimentally vary plausibility of controversial texts by inducing different prior beliefs with short videos while holding prior knowledge constant. We assumed that belief manipulation would affect the perceived plausibility of information in controversial texts. We also assumed that the perceived plausibility, in turn, would influence comprehension in terms of a plausibility effect.

To test this assumption, an experiment was conducted that experimentally varied plausibility of two controversial texts about the medical use of spider silk via short videos. Readers watched one of two versions of a short video on the medical use of spider silk prior to being exposed to two contrary texts discussing the same topic with contrary main claims and competing arguments. The two video versions provided identical factual background information but contained opposing claims on the controversial issue. The pro video argued that spider silk can be used in medicine, whereas the contra video presented the opposite argument. The videos were used to induce a particular belief in participants—that is, either believing in the medical use of spider silk or not—without actually providing evidence or support for this claim. After watching the pro or contra video, participants read two controversial texts on the medical use of spider silk. A verification task was used to assess situation model strength and a validation task was used to assess plausibility judgements. In the validation task, perceived plausibility was assessed for information directly presented in the texts (i.e., paraphrases) and also for information that could be inferred based on the texts' content (i.e., inferences). Hence, perceived plausibility

of the texts was assessed on a finer-grained level, not as global judgment of the plausibility or agreement or belief in a general claim or argumentative stance on the scientific issue.

The experiment created a reading scenario that resembles typical informal learning situations when laypeople search the World Wide Web with the intention to inform themselves about new scientific issues. Readers interested in a new scientific topic often initially come across one-sided information that begins to frame their beliefs before encountering additional information on other web-based sources that are often also one-sided. The conflicting information presented in multiple web-based texts is also not normally directly marked as controversial. Therefore, we did not explicitly alert participants to the controversial nature of the issue in the present study.

In Hypothesis 1, based on the plausibility effect on comprehension (e.g., Maier & Richter, 2013; Schroeder et al., 2008), we expected the comprehension of the text that was consistent with the video version to be better compared to the text inconsistent with the video version. Hypothesis 2 predicted that plausibility judgements would depend on the video version. In detail, participants watching the pro video version would perceive the pro text (i.e., spider silk can be used in medicine) as more plausible compared to the contra text (i.e., spider silk cannot be used in medicine). We expected the reverse would be true for participants watching the contra video version. Most importantly, we expected the effect of the video version on the comprehension of texts to be mediated by readers' plausibility judgements (Hypothesis 3). Thus, we hypothesized and tested a mediation model (Figure 1).

## Method

### Participants and Prior Beliefs

Fifty-nine university students (43 women and 16 men) participated in the experiment. Participants were mainly majoring in Psychology (one participant was majoring in sociology)

with an average semester of 3.08 (*SD* = 1.76) and an average age of 25.41 years (*SD* = 7.25). They received course credit for participation.

Before the main analyses, we investigated the difference score of prior beliefs (agreement to pro belief item – agreement to contra belief item; for details, see *Prior Beliefs* section) to ensure that participants had no strong pre-existing beliefs about the scientific issue before watching the videos. The belief difference scale ranged from –5 to 5 with a theoretical midpoint of 0, with the latter indicating neutrality (e.g. for a participant having the same mean score on the pro and the contra belief scales). Five participants reported extreme prior beliefs as indicated by a difference score of greater or equal 4 or less or equal -4, which deviated strongly from the theoretical midpoint of the scale and hints at a strong initial preference for one argumentative stance in the scientific issue. The data of these participants was not analyzed because belief induction via the video seemed unlikely to work for these participants independent of their pre-existing beliefs.[1] The mean difference score of prior beliefs for the remaining participants was close to zero (*M* = 0.70, *SD* = 1.38). The final sample consisted of 54 participants (41 women and 13 men) with an average of 3.11 (*SD* = 1.81) semesters and an average age of 25.59 years (*SD* = 7.35).

## Materials and Measures

### Prior Beliefs

In the assessment of prior beliefs, participants were provided with 10 divergent statements to five medical scientific topics (e.g., preimplantation genetic diagnosis, stents). This

[1] We performed the analyses with these participants in the data set and by including prior beliefs as covariate and found no relevant change in the reported results.

procedure was used to ensure that participants remained unaware of the controversial medical issue in the focus of the study (i.e., spider silk). In this course, participants' prior beliefs about the scientific issue (whether or not spider silk can be used in human medicine) were assessed with two items (response categories ranging from 1 = *do not agree at all* to 6 = *fully agree*). One item assessed participants' beliefs that spider silk can be used ("I belief that spider silk should be used for a regeneration of nerve tracts in human medicine because this well-tolerated procedure can repair torn nerve structures completely"), and one item assessed participants' beliefs that spider silk should not be used ("I am against the use of spider silk for a regeneration of nerve tracts in human medicine because this expensive and time-consuming procedure can lead to immobility and pain"). A difference score (agreement to pro-belief item – agreement to contra-belief item) was computed for each participant and served as a check to ensure that participants had no strong pre-existing beliefs about the scientific debate.

**Video Material**

To vary the beliefs on the scientific issue, participants watched either a video arguing in favor (pro video version) or against (contra video version) the use of spider silk for a regeneration of nerve tracts in human medicine. The two video versions provided participants with identical background information on spider silk and differed only in seven general statements that took the argumentative stance either for or against the use of spider silk in human medicine (the full text of the videos (translated into English) is available in the OSF repository). For example, the pro version stated that "*Spider silk has many interesting and highly useful characteristics for human medicine*". In the contra version of the video, the same sentence was slightly modified to "*Spider silk has many interesting but not useful characteristics for human medicine*". In all other respect, the two video versions contained the same visual and spoken information and provided the same background knowledge. The videos started with a short introduction and general information

about spiders and their ability to produce spider silk. In the beginning of the video, the uncertain nature of the issue was introduced by stating in both video versions that "*In particular, scientists are trying to clarify whether this type of spider silk can be used to regenerate nerve tissue*." Afterwards, the videos explained the potential use of spider silk in human medicine to repair torn tendons and nerves and discussed existing research findings. Subsequently, more detail was presented on how spider silk can be extracted from spiders. This explanation was followed by a discussion about the production of spider silk in spider glands and of the mechanical properties of spider silk. At the end, the videos made clear that the question of whether spider silk has the potential to be used in human medicine has been unanimously decided and ended with a concluding statement that was in line with the video version. In the pro version this statement concluded, "*Intensive empirical investigations from well-respected research institutes have shown compellingly that spider silk can be used to repair torn nerves and tendons in human medicine,*" whereas the contra version concluded, "*Intensive empirical investigations from well-respected research institutes have shown compellingly that spider silk cannot be used to repair torn nerves and tendons in human medicine.*" The total length of the video was 7.42 minutes. The two video versions were pretested with an independent sample of university students ($N = 61$) for potential differences. Results revealed that both the pro and the contra version of the video were perceived as similar in critical aspects such as understandability, plausibility, and interest (Table 1). In addition, we found no difference in general judgments about the video such as quality of sound and picture.

**Text Material**

Two texts about the scientific question of whether or not spider silk has the medical potential to repair torn tendons and nerves were used as experimental material. We selected this scientific debate for the experimental texts because an independent sample of university students

(*N* = 38) indicated in a pilot study with 25 different scientific topics that they possessed almost no prior knowledge about this issue (*M* = 1.08, *SD* = 0.27; ratings on a scale from 1 = *no prior knowledge* to 6 = *high amount of prior knowledge*), had not read or heard about this issue before (*M* = 1.08, *SD* = 0.36; ratings on a scale from 1 = *no exposure* to 6 = *high amount of exposure*), but at the same time were interested in reading more information about the scientific issue (*M* = 3.11, *SD* = 1.52; ratings on a scale from 1 = *no interest* to 6 = *high amount of interest*). Moreover, the difference in the mean agreement to the two argumentative positions in this controversy (i.e., "Spider silk has the medical potential to repair torn tendons and nerves" vs. "Spider silk does not have the medical potential to repair torn tendons and nerves"; agreement to each item indicated on a rating scale from 1 = *do not agree at all* to 6 = *fully agree*) was close to zero in this pretest (*M* = -0.11, *SD* = 1.20) indicating that there was no clear preference for one argumentative position in this scientific debate. This lack of preference was a necessary precondition to ensure that differences in the perceived plausibility of the text material would be due to the belief effects induced by the video and would not be due to effects of pre-existing beliefs.

The two experimental texts on spider silk took opposing positions in the scientific controversy, were comparable in length and readability, and followed the same rhetorical structure (Table 2). The *pro text* argued that spider silk is able to repair torn nerves and should be used in surgery. In contrast, the *contra text* argued against the use of spider silk in surgery. Each text presented three unique arguments separated by subheadings that consisted of a claim that was followed by supporting evidence. The arguments were always supportive for the main claim of the text. The texts were pretested with an independent sample of university students (*N* = 12) to ensure that the texts did not differ in their characteristics (understandability, argument quality, plausibility, and clearness of the stance toward the issue, see Table 1). Wilcoxon signed-rank test

revealed no significant differences between the texts in the text characteristics. Moreover, participants in the pretest were able to successfully indicate the argumentative stance of the texts, (Table 2) and a significant difference between the texts was found on participant's judgements of the position of each text (response category ranging from 1= *spider silk should not be used* to 7 = *spider silk should be used)* in the Wilcoxon signed-rank test, $p < .001$.

**Comprehension Measure**

Text comprehension was measured with 24 test items (sentences) per text with a verification task (modified after Schmalhofer & Glavanov, 1986). Test items were either paraphrases of the text information, inferences that matched the content of the text, or distracters (eight items per item type per text). Table 3 provides an example for each item type for the pro and the contra text. Paraphrases were created by varying the word order of a sentence from the text and replacing key content words with synonyms. Inferences contained information that was not explicitly stated in the text but instead needed to be inferred by the participants to build an adequate referential representation of the text. Finally, distracters communicated information that was neither mentioned explicitly in the text nor a sensible inference from the text but shared some superficial content aspects with the text. In the verification task, participants indicated whether or not each test sentence can be inferred from the text. The accuracy of the responses to the paraphrase and inference items was investigated as an indicator of comprehension, and responses to distracter items in the verification task were used to estimate participants' response bias.

**Plausibility Judgements**

For each test item used in the verification task, participants provided binary plausibility judgments (plausible vs. implausible). They were instructed to answer "yes, sentence is plausible" if they thought that the information presented in the test sentence is (presumably) true.

Plausibility judgements to the paraphrase and inferences items were investigated as an indicator of perceived plausibility, and responses to distracter items in the validation task were again used to estimate participants' response bias.

**Manipulation Checks**

In a first manipulation check, participants were asked to indicate the argumentative stance of the video that they had watched with two items. One item stated that the video took the pro position in the controversy ("*The video argued for the use of spider silk in human medicine*") and one item stated that the video took the contra position in the controversy ("*The video argued against the use of spider silk in human medicine*"). Responses were provided on a scale ranging from 1 = *fully incorrect* to 7 = *fully correct*.

A second manipulation check examined whether participants noticed that the texts were conflicting. For this aim, one item ("*The texts took the same argumentative stances*") with a response category ranging from 1 = *do not agree at all* to 7 = *fully agree*) was used.

**Procedure**

At the beginning of the experiment, participants' prior beliefs were assessed. Participants then watched either the pro or contra video on the medical use of spider silk and subsequently read the two texts about the scientific issue of the medical use of spider silk in a self-paced fashion on a computer screen. After reading both texts, participants provided responses to the test sentences, which were presented one-by-one in black letters (font type Arial, average height 0.56 cm, bold) on a white background and in random order. Participants indicated by pressing one of two response keys marked green (*yes*) and red (*no*) in the verification task to indicate whether or not the test sentence could be inferred from the texts. In addition, participants provided plausibility judgments for the same set of test items in the validation task. The order of the verification task and the validation task was varied between participants. After working on the

verification and validation task, participants were asked to indicate the argumentative stance of the video that they had watched with two items as a manipulation check. At the end of the experiment proper, participants were thanked and debriefed.

### Design

The core experimental design was a 2 (*video version*: pro vs. con, varied between participants) x 2 (*text type*: pro vs. con, varied within participants) mixed design. In addition, *text order* (pro-con vs. con-pro, varied between participants) and *task order* (verification – validation vs. validation – verification, varied between participants) were counterbalanced between participants. Participants' response bias (assessed with participants' response to the distracters) was included as a covariate in the analyses.

## Results

### Data Analysis Strategy

Following the traditional causal steps approach to mediation analysis (Baron & Kenny, 1986), four subsequent steps were taken to investigate our hypotheses. In Step 1, we tested the direct effects of the predictors *video version* and *text type* on comprehension as the outcome variable (i.e., accuracy of responses to paraphrases and inferences in the verification task). The interaction is crucial for the effect predicted in Hypothesis 1. In Step 2, we tested the effect of the distal predictors *video version* and *text type* and their interaction on the potential mediator *plausibility*. The interaction term in this model is crucial for the effect predicted in Hypothesis 2. In Step 3, we tested the effect of the potential mediator *plausibility* on comprehension while controlling for the effects of the distal predictors *video version* and *text type*. In Step 4, we examined how the direct effect of the predictors on comprehension changes after the mediator is included in the model. If Hypothesis 3 holds, the interaction of video version and text type will

become statistically nonsignificant. Finally, in addition to the causal steps approach, we also estimated and tested the indirect effect predicted in Hypothesis 3.

The data in the present study has a multilevel structure. Plausibility and verification judgements were assessed on the item level, that is, as responses (correct/plausible vs. incorrect/implausible) to paraphrases and inferences. In addition, the experimental manipulation occurred on the subject level with half of the participants each watching one of the two video versions for belief induction. In analyzing the data, we accounted for the multilevel structure of the data by estimating generalized linear mixed models (GLMM) with the logit link function. We used the lme4 package with bobyqa as the optimizing function (Version 1.1-21, Bates et al., 2015) and the lmerTest package (Version 3.1-0, Kutznetsova et al., 2014) in R (Version 3.6.0., R Core Team, 2019). Type I error probability for all significance tests was set to .05. The models specified the fixed effects of the contrast-coded independent variables *text type* (-1 = contra text, 1 = pro text), *video version* (-1 = contra video, 1 = pro video), the *order of the texts* (-1 = contra text first, 1 = pro text first), the *order of the tasks* (validation task first = -1, verification task first = 1) and their interactions as predictors. In Step 3, the fixed effect of *perceived plausibility* (-1 = implausible, 1 = plausible) was included as potential mediator. The order of the texts and the order of the tasks were entered to control for ordering effects and to account for the experimental design of the study. In addition, participants' response to the distracters (grand-mean centered) were included in all GLMM models to account for response biases such as guessing or a general tendency to provide "yes" responses. In Step 1, the accuracy of responses to the paraphrase and inference items in the verification task was used as dependent variable to investigate the effects of the distal predictors *video version* and *text type* on comprehension. In Step 2, the plausibility judgments to the paraphrase and inference items served as the outcome variable to investigate the effects of the distal predictors *video version* and *text type* on the potential mediator *perceived*

*plausibility*. In Step 3, the accuracy of responses to the paraphrase and inference items in the verification task was used as the dependent variable and the fixed effects of *video version*, *text type*, and the effect of the potential mediator *perceived plausibility* were analyzed. In Step 4, the indirect effect was tested in a design with a subject-level treatment (i.e., video version), and a test item-level mediator (i.e., plausibility) and outcome variable (i.e., verification responses). The mediation package (Version: 4.5.0, Tingley, et al., 2019) for R was used to estimate causal mediation effects in our multilevel data.

Effect sizes (Cohen's $f^2$) for significant fixed effects were determined with a model-comparison approach. We compared the variance explained by the model without the focal effect to the full model that included the effect. The variance explained by the fixed and the random effects in each model was determined using the R-package MuMIn (Version: 1.43.17, Bartoń, 2020). Observed power analyses were conducted with the R-package simr (Version: 1.0.5, Green & MacLeod, 2016) that allows for power analysis of GLMM by simulation. To estimate the observed power (1-β) for independent variables (e.g., for the interaction of video version and text type and for plausibility), 1,000 simulations based on the design and sample size of the experiment were used.

We report only significance tests relevant to the hypotheses in the text, that is, main fixed effects of the independent variables *video version* and *text type* as well as their interaction and main fixed effects of the potential mediator *perceived plausibility* in Steps 3 and 4. Descriptive statistics of all variables are provided in Table 4 and parameter estimates for the three GLMMs are provided in Table 5. In addition to parameter estimates, we report predicted (conditional) probabilities of the responses (back-transformed from the logit-link model with estimated standard errors).

In additional analyses, we scrutinized the proposed mediation model further by comparing it to a number of alternative models. To this end, we specified and compared the fit of four nested models with comprehension as dependent variable:

Model 1a) was a null model containing no fixed effects, but the random effects for subjects and test items,

Model 2a) was a baseline model with the control variables (added as fixed effects) reading order (-1 = contra text first, 1 = pro text first), task order (validation task first = -1, verification task first = 1), the interaction of reading order and task order as well as participants' responses to distracters (grand-mean centered),

Model 3a) additionally included the fixed effects of text type (-1 = contra text, 1 = pro text), video version (-1 = contra video version, 1 = pro video version), their interaction and the interactions of the between-subject factors and model (Model 3a was identical to the unmediated model tested in Step 1),

Model 4a) additionally included the fixed effect of plausibility (-1 = implausible, 1 = plausible) as potential mediator (Model 4a was identical to the mediated model tested in Step 3).

To allow a comparison of these four models to a reverse mediation model, in which plausibility is viewed as dependent variable and comprehension response as potential mediator, four similar models were specified for plausibility as dependent variable.[2] In this case, the same random and/or fixed effects were entered in the null model (Model 1b), the baseline model (Model 2b) and the unmediated model (Model 3b). Model 4b was a reverse mediation model with

[2] We thank an anonymous reviewer for highlighting the importance of model comparison and the need to test the possibility that a reverse mediation model might fit the data equally well.

plausibility as outcome and the fixed effect of comprehension accuracy (-1 = incorrect answer, 1 = correct answer) as potential mediator.

Several indices of model fit (log-likelihood, AIC, BIC, deviance) that allow an evaluation of the goodness-of-fit of one model in relation to other models are reported in Table 6. Table 6 also provides likelihood-ratio tests for nested models. The proposed mediation model (Model 4a) and the reverse mediation model (Model 4b) are not nested. Therefore, the comparison of these two models relies on a descriptive comparison of the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) with the assumption that lower scores indicate better fit for different models of one model class. We further provide Pseudo-$R^2$ for the unmediated Models 3a and 3b and the mediated Models 4a and 4b to provide additional information on model fit. Pseudo-$R^2$ may be construed as the variance explained by the entire mixed-effects models, including both fixed and random effects and was determined using the R-package MuMIn (Version: 1.43.17, Bartoń, 2020).

**Data Cleaning**

Prior to the main analyses, the data set was checked for outliers. The data analysis strategy is a generalization of regression analysis that might be influenced by outliers or violations of the assumptions underlying regression analysis (cf. Cohen, Cohen, West, & Aiken., 2003). In detail, the R package influence.ME (Version: 0.9.9, Nieuwenhuis, Grotenhuis, & Pelzer, 2012) was used to compute Cook's *D* (Cook, 1977) and $DFBETAS_{ij}$ for each model. The R-package influence.ME provides these indicators for mixed effects models estimated with lme4 by accounting for the nested structure of the data. Cook's *D* is an indicator of the combined effect of leverage (extremity in the independent variables) and discrepancy (extremity in the dependent variables). DFBETASij is a local measure of influence, that is, it captures the influence of individual data points on specific regression coefficients. On both indicators, higher values

indicate a larger influence. Following the suggestions of Van der Meer, Te Grotenhuis, and Pelzer (2010), the cut-off value for Cook's *D* was set to 4/*n* with *n* being the number of Level 2 groups (in our case participants, leading to 4/54 = 0.07). Note that this value is below the cut-off value suggested by Cohen et al. (2003) of 1.0. For DFBETASij, the cut-off value was set to $2/\sqrt{n}$ ($2/\sqrt{54} = 0.27$), which is again below the cut-off value for DFBETASij of ±1 suggested by Cohen and colleagues. One participant slightly exceeded the cut-off values multiple times, but simulation data suggested no change in significance of relevant effects from the inclusion of this participant. Hence, data from all participants were included in the analysis.[3]

Furthermore, we scrutinized individual data points that seemed unlikely to be based on valid responses by participants or that were caused by extreme implausibility of the test items as a consequence of material construction. In detail, the number of possible observations available in our study was 1,728 (54 participants x 32 items). Given that plausibility was the main predictor in our study and was expected to depend on the video version as experimental manipulation, we first checked whether all test items were globally plausible (i.e., no rating of high implausibility). After inspection of the global plausibility ratings for the test items, we found that two test items were considered highly implausible by all participants as indicated by a mean plausibility of .50. Consequently, these two items were removed from further analysis, leading to a number of possible observations of 1,620 (54 participants x 30 items). From these 1,620 observations, individual responses to test items that were below 100 ms were excluded because it seemed highly unlikely that participants were able to read and respond to the test items within this time

[3] We performed the analyses without this participant in the data set and found no relevant change to the reported results.

frame. Moreover, responses to individual test items with response times deviating more than two standard deviations from the mean of the test item were also not analyzed because this length of time indicates that participants were not fully reading the item or had been distracted when responding (which might result, e.g., in mistakenly pressing a response key). The remaining number of observations in this study was 1,541 and all responses to the test items that were excluded based on the criterions were treated as missing values. Hence, 79 (5%) of the 1,620 were missing values. We found no indication that the resulting pattern of missing values was systematic in any way. The GLMM can be estimated even if single values of a participant are missing in the dataset.

**Manipulation Checks**

To compare the two manipulation check items, we computed an ANOVA with between- and within-subjects factors. We found a significant interaction of video version and item stance, $F(1,46) = 122.46$, $p < .05$, $\eta_p = .73$. Participants who had watched the pro video version agreed more strongly that the video argued for the use of spider silk ($M = 6.64$, $SE = 0.28$) and less that the video argued against the use of spider silk ($M = 1.22$, $SE = 0.30$), $F(1,46) = 98.03$, $p < .05$, $\eta_p = .68$. Participants who had watched the contra video version correctly agreed more strongly that the video argued against the use of spider silk ($M = 5.63$, $SE = 0.31$) and less that the video argued for the use of spider silk ($M = 2.32$, $SE = 0.29$), $F(1,46) = 33.99$, $p < .05$, $\eta_p = .42$. The results indicated that we could be certain participants would discern the stance taken in the videos and that the central manipulation had the intended effect.

In addition, we investigated whether participants could discern that the texts presented opposing viewpoints. Mean agreement to this manipulation check item was 1.17 ($SD = 0.51$),

demonstrating that participants could discern that the texts took opposing viewpoints on the scientific issue.

**Step 1: Effects of video version and text type on the accuracy in the verification task**

The effects of the main predictors in the unmediated model are illustrated in Figure 2. In the first GLMM with accuracy in the verification task as dependent variable, we found no main effect of text type (β = 0.04, $z$ = 0.26, $p$ = .79) or video version (β = -0.15, $z$ = -1.07, $p$ = .29). However, as predicted by Hypothesis 1, we found a significant interaction of text type and video version (β = 0.19, $z$ = 2.80, $p$ < .05, Cohen's $f^2$ = .01). Readers who watched the contra video version provided more accurate responses to inference test items from the contra text ($P$ = .86, $SE$ = .03) compared to readers who watched the pro video version ($P$ = .76, $SE$ = .05), β = -0.33, $z$ = 2.16, $p$ < .05. Readers who watched the pro video version, however, provided no more accurate responses to the test items from the pro text ($P$ = .83, $SE$ = .03) compared to readers that watched the contra video version ($P$ = .82, $SE$ = .04), β = 0.04, $z$ = 0.28, $p$ = .79. In sum, this response pattern is partly in line with Hypothesis 1. The observed power for detecting the focal interaction of video version and text type in the unmediated model was .81 (95% CI [0.78, 0.83]).

**Step 2: Effects of video version and text type on plausibility judgments**

In the GLMM model for the plausibility judgments as dependent variable, we also found no main effects of text type (β = -0.00, $z$ = -0.01, $p$ = .99) and video version (β = 0.04, $z$ = 0.38, $p$ = .71), but an interaction of text type and video version (β = 0.30, $z$ = 4.88, $p$ < .05, Cohen's $f^2$ = .03) as predicted by Hypothesis 2. Readers who watched the contra video version perceived the contra test items ($P$ = .80, $SE$ = .04) as more plausible compared to readers who watched the pro video version ($P$ = .71, $SE$ = .05), β = -0.26, $z$ = -2.05, $p$ < .05. Similarly, the pro test items were judged as more plausible by readers who watched the pro video version ($P$ = .81, $SE$ = .03) compared to readers who watched the contra video version ($P$ = .69, $SE$ = .05), β = 0.34, $z$ = 2.79,

$p < .05$. This pattern of results is fully in line with Hypothesis 2 (see Figure 3). The observed power for detecting the focal interaction of video version and text type on the potential mediator plausibility was .99 (95% CI [0.99, 1.00]).

**Step 3: Effects of video version, text type and plausibility on comprehension**

The third GLMM with accuracy in the verification task as dependent variable investigated the effect of the potential mediator perceived plausibility on comprehension while the effects of the distal predictors text type, video version, and their interaction were statistically controlled. Again, we found no main effect of text type ($\beta = 0.05$, $z = 0.43$, $p = .67$) or video version ($\beta = -0.17$, $z = -1.36$, $p = .17$). However, we found a main effect of plausibility ($\beta = 0.95$, $z = 12.26$, $p < .05$, Cohen's $f^2 = .12$), but the interaction of text type and video version was no longer significant ($\beta = 0.07$, $z = 1.03$, $p = .30$). This pattern of results is in line with the mediation predicted in Hypothesis 3 and depicted in Figure 4. The observed power for detecting the effect of plausibility in the mediated model was 1.00 (95% CI [0.99, 1.00]).

**Step 4: Test of indirect effect**

In Step 4, we computed the estimated mediation, direct, and total effects to investigate the significance of the indirect effect. The average causal mediation effect (Estimate = 0.02, 95% CI Lower = 0.00, 95% CI Upper = 0.03, $p = .008$) of video version and text type on comprehension via perceived plausibility and the average total effect (Estimate = 0.03, 95% CI Lower = 0.01, 95% CI Upper = 0.05, $p = .018$) were significantly different from zero. In contrast, the average direct effect of the interaction of video version and text type was not significant from zero (Estimate = 0.01, 95% CI Lower = -0.01, 95% CI Upper = 0.03, $p = .29$). In sum, the results support our mediation model and suggest that the belief manipulation, which was experimentally varied between subjects with the video version, affected the perceived plausibility of paraphrases and inferences from the two texts differently, which in turn influenced the comprehension of the

two texts differently. In detail, the text that was consistent with the video version had a higher likelihood of being perceived as plausible, and this higher plausibility perception increased the comprehension for this text, that is, the amount of correct responses to inferences and paraphrases. It is important to note that the effect of the video version on comprehension of the two controversial texts was fully mediated by perceived plausibility given that the direct effect was not significant after including the mediator in the model.

**Model Comparison**

To investigate whether the reported mediation model is the model with the best fit to the data as well as the model that explains most of the variance of the entire mixed-effects models, we compared the nested models for comprehension as outcome (see Table 6, upper part). Descriptive statistics as well as the likelihood ratio test suggested the best model fit for the proposed mediation model for comprehension (Model 4a). In addition, a similar pattern was found for plausibility as outcome variable (see Table 6, lower part).[4] A comparison of the proposed mediation model (Model 4a) with the reverse mediation model (Model 4b) based on the AIC and BIC scores revealed that the proposed mediation model had a better fit to the data compared to the reverse mediation model. In addition, Pseudo-$R^2$ was higher for the proposed mediation model (Model 4a: Pseudo-$R^2$ = .37) compared to the reverse mediation model (Model 4b: Pseudo-$R^2$ = .29). For the unmediated model, we found similar effects. Again, Pseudo-$R^2$ was

[4] In addition, in the reverse model with plausibility as outcome, the average causal mediation effect was not significant (Estimate = 0.01, 95% CI Lower = -0.003, 95% CI Upper = 0.02, $p$ = .14). The average direct effect of the interaction of video version and text type on plausibility as dependent variable was significant (Estimate = 0.04, 95% CI Lower = 0.02, 95% CI Upper = 0.06, $p$ < .05) indicating no mediation on plausibility judgments via comprehension.

higher for the model with comprehension as outcome (Model 3a: Pseudo-$R^2$ =.30) compared to the model with plausibility as outcome (Model 3b: Pseudo-$R^2$ =.22). Together, these results allow for the conclusion that the proposed mediation model with perceived plausibility fully mediating the effect on comprehension can be considered the model that fits the data best.

## Discussion

The present study investigated whether a belief manipulation (i.e., watching a video taking a pro or con stance) would influence the perceived plausibility of information from an unfamiliar controversial debate, which in turn, would lead to a plausibility effect on comprehension (i.e., a better comprehension of plausible information). Results were in line with the hypothesized mediation model. The indirect effect of the interaction of video version and text type on comprehension via perceived plausibility was significant, whereas the direct effect of the interaction of video version and text type (which was significant in the unmediated model) was no longer significant when perceived plausibility was included in the mediation model. In other words, the better comprehension of the text that took the same argumentative stance as the video that the participants had watched was fully mediated by perceived plausibility. This finding is remarkable because it suggests that readers' comprehension of scientific information may suffer when it is inconsistent with their initial understanding. The perceived plausibility might be one factor that could hinder the understanding of alternative arguments and explanations, which is an important pre-requisite for revising one's standpoint or even for conceptual change. In other words, if readers do not understand or comprehend a counter-argument in a scientific debate, they will not able to rationally decide whether their mental model or standpoint on the issues needs to be revised.

The results are in line with earlier research indicating a significant role of plausibility in comprehension (e.g., Maier & Richter, 2013; Schroeder et al., 2008). However, earlier research

had been conducted with logical errors and fallacies to objectively vary plausibility or did not experimentally manipulate plausibility. The present study is a major step forward because the experimental procedure allowed us to assess plausibility as subjective judgements that were nevertheless experimentally varied by inducing different prior beliefs with short videos. Hence, the results of this study provide strong evidence for the causal link between perceived plausibility and comprehension. Of course, for a better generalization, it will be important to replicate the findings with different samples and different scientific topics.

The results of the present study also shed some light on the type of information readers can use as epistemic background for plausibility judgements. The present study used two videos versions (pro vs. con) to vary perceived plausibility of the controversial texts. Apart from arguing for divergent positions in the controversy on spider silk, the two versions of the video provided the same background information on the scientific debate. Using informationally equivalent versions allowed us to rule out the possibility that differences in comprehending the two texts are caused by differences in prior knowledge. Lombardi and colleagues (2016), for example, argued that scientists and lay people have different plausibility perceptions because of their different expertise or level of prior knowledge. In detail, they assume that a plausibility gap exists between scientists and lay people, which often hampers conceptual change for lay people, that is, replacing false knowledge with new correct knowledge. Results from our study suggest that in some circumstances, mere beliefs are used as epistemic background for plausibility judgements during comprehension. Both videos used as experimental manipulation provided identical factual information on the scientific issue of spider silk. However, the videos were varied in such a way that one argumentative side (pro vs. contra) on the scientific issue of spider silk was depicted as the correct one. Still, we found that readers who watched the contra video version perceived the information from the contra text as more plausible, whereas readers who watched the pro video

version perceived information from the pro text as more plausible. Plausibility was then directly connected to comprehension. This finding is important because it shows that readers might simply assess how well a new piece of information coheres with their (even experimentally induced) beliefs about a scientific controversy.

The effects of the video version as belief manipulation on plausibility and on comprehension are also fully in line with the predictions made in the Two-Step Model of Validation (Richter & Maier, 2017). This model assumes that readers often tend to allocate their cognitive resources to information judged as plausible based on non-strategic validation. In the present study, no processing or behavioral data was assessed as the focus of the study was on the effects of experimental induced beliefs and plausibility on the comprehension outcomes. Nevertheless, the effects of the video version on plausibility might be interpreted in the light of the Two-Step model of Validation and its assumption that readers might have used the experimentally induced beliefs for non-strategic validation. Moreover, the effect of plausibility is in line with the idea that plausibility judgments, according to theory and research on validation, occur regularly and continuously during reading and lead to better comprehension for such information when no strategic attempts are made to resolve inconsistencies or contradictions (Richter & Maier, 2017).

Using the perceived plausibility of information as a heuristic to comprehend conflicting texts seems to be an easy way to maintain a coherent mental representation of a controversial issue. Nevertheless, an attached detriment of such heuristic processing is that readers seem to primarily understand information that is implicitly judged as plausible. In our experimental procedure, this is not necessarily the information that can could easily be positively evaluated with background knowledge but instead only the information that was consistent with an experimentally induced belief. A similar mechanism might be at work when readers fail to

comprehend new information about a topic that is at odds with earlier information (e.g., Blanc, Kendeou, van den Broek, & Brouillet, 2008; Johnson & Seifert, 1994) or information from belief-inconsistent texts (e.g., Abendroth & Richter, 2020a, Maier & Richter, 2013). Given that the strong and sustainable effects of beliefs as a special kind of gatekeeper that allocates resources for comprehension seem to occur in many reading situations (for an overview, see Richter & Maier, 2017), identifying educational interventions that are able to reduce the impact of prior beliefs on comprehension is an important educational objective. Promising avenues for successful interventions seem to be specific reading goals (e.g., Bohn-Gettler & McCrudden, 2018; Maier & Richter, 2016; Wiley & Voss, 1999), an interleaved presentation of controversial texts (Abendroth & Richter, 2020a; Maier & Richter, 2013; Wiley, 2005), and metacognitive strategy trainings (Abendroth & Richter, 2020b; Maier & Richter, 2014). For example, Bohn-Gettler and McCrudden (2018) found a positive effect of relevance-task instructions on the recall of a belief-relevant dual-position text to the extent that no belief-effects on recall occurred. Participants were either instructed to focus on pro or on contra information while reading a dual-position text. This task-relevance instruction influenced recall to the extent that task-relevant text was better recalled than task-irrelevant text – independent of readers' prior beliefs. Note that in this study participants' beliefs still impacted the strategic processing of the texts because think-alouds indicated more confirmation strategies for belief-consistent text and more disconfirmation strategies for belief-inconsistent text – independent of the task instruction.

The focus of the present study tested the effects of perceived plausibility, varied by experimentally-induced prior beliefs, on the comprehension of a controversy. Readers' judgements about texts (e.g., plausibility or truth judgements) might, however, also be influenced by different heuristics such as cognitive ease of processing, fluency, or truth effects based on

repetition (Brashier & Marsh, 2020).[5] For example, repeatedly being exposed to the same factual information, such as hearing or reading the same sentence or message more than once, increases its fluency, familiarity, and recognition (Unkelbach, Koch, Silva & Garcia-Marques, 2019). This alternative interpretation seems less likely for the present results. Both video versions contained the same factual information, that is, all participants received the same information before reading the texts. In addition, the texts provided arguments on the scientific debate that were not addressed in the videos. Hence, no factual information from the videos was repeated by the texts. Arguably, the consistency between the general stance on the scientific issue presented in the video version and the text that took the same stance could be interpreted as repeating factual information. However, the verification task used to assess comprehension did not focus on the general stance of the video or texts but instead assessed comprehension on a finer-grained level using paraphrases and inferences from the texts. We investigated reading times for the two texts to additionally rule out the alternative interpretation that the ease of cognitive processing or fluency, due to repetition, caused the investigated difference in plausibility and memory of the two texts. Cognitive ease or fluency is characterized by little cognitive load on information processing and has direct consequences on reading times. Following this assumption, the text that was consistent with the video version in our experimental scenario should have been easier to process because less memory interferences occur during reading this type of text. An ANOVA with between- and within-subjects factors found no interaction of video version and text type on reading times, $F(1,46) < 1.0$, n.s., indicating that reading times for the two texts did not vary as a function of the video version. This finding gives good reason to assume that cognitive ease or

[5] We thank two anonymous reviewers for highlighting this alternative interpretation.

fluency was not the driving force in our experiment but rather that differences in plausibility judgments and comprehension can be attributed to the experimentally induced beliefs.

**Plausibility Judgements, Conceptual Change and Belief Revision**

The present study investigated how plausibility judgements influence the comprehension of multiple texts on scientific debates. Such results are likely to have consequences for processes that depend on understanding arguments and counterarguments such as conceptual change or belief revision. For example, Posner, Strike, Hewson and Gertzog (1982) expected that conceptual change can occur when a new conception or alternative explanation for a phenomenon or central concept appears plausible. Posner and colleagues further assumed that plausibility in turn is influenced in five ways, which all have in common that they depend on strategic considerations and can be a key for accommodation processes. This assumption means that when a new alternative explanation is strategically judged as plausible, it might lead to a reorganization or replacement of readers' central concepts on the issue. Lombardi et al. (2016) similarly argued that plausibility judgements might be one crucial element in conceptual change and that more explicit plausibility judgements are needed to reappraise more implicit plausibility judgments. The plausibility effect investigated in the present study seems not to be a possible source for accommodation – or conceptual change – but instead one factor that might hinder accommodation or situation-model updating. This plausibility effect was observable with the condition of participants having a better comprehension of plausible information and a weaker understanding for information perceived as implausible. In our study, the plausibility judgements were experimentally varied by inducing different beliefs via a short video. As a consequence, readers understood the text that took the opposing view of the video to a lesser extent than the text that argued for the same position in the controversy. For conceptual change to occur, understanding the alternative explanation is a crucial prerequisite: If readers do not understand or

comprehend the counter-argument, they are unlikely to be able to critically decide whether their mental model needs to be revised. Hence, the plausibility effect found in the present study could have been due to implicit plausibility judgments which might have been a source of assimilation (by simply paying less attention to information judged as implausible). In these circumstances, situation model updating or conceptual change is likely to occur when epistemic elaboration processes reduce the impact of prior beliefs or implicit plausibility judgments (Richter & Maier, 2017).

Situation model updating or conceptual change might also include processes of belief revision. The theory of explanatory coherence (Thagard & Findlay, 2010) addresses the question under which circumstances belief revision occurs. The theory describes that belief revision is often proceeded by evaluating all relevant alternatives with all available arguments and evidence, which is a rather strategic evaluation process that requires that arguments and counterarguments for a topic are fully understood. This theory does not focus on comprehension or memory for information but instead makes predictions for the construction and revision of a belief system. Accordingly, we think that the results reported in the present study are relevant to explain why people often refrain from adopting new (or more correct) views on an issue. That is, beliefs or plausibility judgements affect what readers understand when reading controversial information. Given that beliefs were experimentally varied in our study, “hot coherence” was unlikely to play a role in not understanding information judged as implausible. Instead, our results showed that plausibility judgments can hamper conceptual change or belief revision because implausible information is simply understood to a lesser extent.

**Limitations and Implications for Future Research**

Plausibility judgements can occur on different levels ranging from judging the semantic plausibility of sentences based on concept or word coherence (e.g., Connell & Keane, 2004) to

judging the plausibility of an entire situation or event based on contextual cues (e.g., Rapp, Hinze, Slaten, & Horton, 2013). The judgments might further be based on the texts' content (Schroeder, Richter, & Hoever, 2008) or on source information (de Pereyra et al., 2014). In the present study, we were interested in how prior beliefs affect the perceived plausibility and by that comprehension. To examine this question, readers provided plausibility judgments for information directly presented in the texts (i.e., paraphrases) and also for information matching the situation described in the text (i.e., inferences). However, perceived plausibility might not only vary as a function of the degree of consistency between text information and readers' beliefs and knowledge but might be additionally impacted by the credibility of the information source (Wertgen & Richter, 2020). In their study, Wertgen and Richter systematically varied perceived plausibility of short stories by inserting true or false factual information, and they also systematically varied the credibility of the source (i.e., either high or low credibility based on expertise). In two experiments, an interaction of plausibility and source credibility for implicit and explicit measures of validation was found. For future studies it would therefore be interesting to investigate how plausibility judgements based on prior beliefs are affected by source features or additional features of the text and the reading situation. One possible focus could be, similar to Wertgen and Richter (2020), the interaction of source credibility and plausibility judgements that are based on prior beliefs. This interaction seems likely in a situation in which source credibility is a salient characteristic and can be easily assessed by readers. If source information is not a salient characteristic or easily assessed, beliefs might not only influence the perceived plausibility of controversial argumentative texts about an unfamiliar topic but might also influence the perceived plausibility or credibility of the text sources. This outcome would suggest the possibility that plausibility judgements could be extended to the judgement of sources. Such an idea would complement existing theories on source judgements, which often suggest that source

or credibility features are used to evaluate text information (e.g., Bråten, Strømsø, & Britt, 2009). For example, the Discrepancy-Induced Source Comprehension (D-ISC) Model (Braasch, & Bråten, 2017) proposes that conflicting messages stimulate readers to use source information to evaluate the trustworthiness of claims from controversial texts. Such a process is viewed as strategic and is likely to be resource intensive. Based on the notion of nonstrategic effects of plausibility during validation, judging sources from plausible or belief-consistent texts as more trustworthy or credible also seems likely. Future studies should empirically test this possibility.

Plausibility judgements are often viewed as a matter of degree (e.g., Lombardi et al., 2016). In the present study, a dichotomous assessment of plausibility judgements was used at the item level to enable participants to respond as fast as possible and without much strategic thought or consideration to each item in the validation task. Using a rating scale, for example, to assess plausibility at the item level would have led to a more strategic evaluation of the test items' plausibility, but such strategic evaluation was not the focus of the current study. Nevertheless, the resulting plausibility scores vary as a matter of degree at the text level. Future research should test the effectiveness of alternative procedures to assess plausibility on a fine-grained level but without much conscious thought of the participants, for example, by using three-answer options (plausible, neutral, implausible).

In the present study, a video was used to vary participants' beliefs because the aim was to present relevant background information on spiders, spider silk, and its medical use in an easy way by describing relevant processes in spoken words combined with dynamic pictures (e.g., on how spiders extract spider silk and weave their net). The multimedia principle suggests that readers are able to learn better from the combination of (spoken) words and (dynamic) pictures (Mayer, 2017). In our experimental scenario using a video to provide the introduction into the scientific topic (in contrast to, e.g., a third text) might thus have been beneficial for learning the

basic facts about spider silk and its' medical use. In addition, the video allowed for a fairly long introduction into the topic with only seven sentences differing between the two video versions. The role of multimedia vs. text-only presentations for forming beliefs that affect the processing of later information was not the focus of the present study. However, one potentially interesting avenue for future research is an investigation of the plausibility effect on the comprehension of scientific debates when participants are confronted with different combinations of text, text and pictures, and videos, which is typical for informal learning on the Web.

We selected the scientific issue of the medical use of spider silk for the experimental materials because participants had no clear preference for one argumentative stance in the debate. They also possessed almost no prior knowledge but were interested in reading more about the issue. This choice of topic allowed us to experimentally vary participants' beliefs. For this experimental manipulation, only seven sentences differed between the two video versions, which can be considered a rather minimalistic experimental manipulation. In fact, only small effects of the interaction of video version and text stance and small-to-medium effects of plausibility were found. The medium effect size for plausibility on the dependent variable is in accordance with other research on plausibility. For example, Körner, Joffe and Deutsch (2019) found a comparable medium effect size for plausibility on moral dilemma judgment scores. A stronger link between beliefs and plausibility might occur for scientific topics for which beliefs are not experimentally varied but are based on a longer learning history. In such topics, stronger beliefs might be created because a scientific controversy is personally relevant for a reader, such as the risks and consequences of vaccines for a mother of a newborn. If prior beliefs are strong, we would expect that readers use these beliefs to validate new textual information more often and more efficiently. As a consequence, the link between belief-consistency and plausibility might be stronger for such readers. Whereas there is growing research on the influence of (strong) prior

beliefs on text comprehension (for an overview see Richter & Maier, 2017), to the best of our knowledge no research to date investigated the effect of strong prior beliefs on the role of plausibility in comprehension. Future research on scientific debates with a clear preexisting preference of readers for one argumentative stance should be conducted to further investigate the link between beliefs, plausibility, and comprehension. This might also be complemented by comparing participants with strong prior beliefs with participants with weak prior beliefs.

In addition to investigating the role of plausibility for scientific topics for which readers have strong prior beliefs in future research, it would further be important to investigate whether the effects found in the present study also apply to other scientific issues and a wider population. Replicating the findings with different samples and different scientific topics would be important for a better understanding of the role of plausibility in comprehension.

The comprehension measure used in this study targeted the individual texts and not intertextual comprehension or integration between the two texts. Future research targeting across-text integration as the dependent variable would augment our understanding of the role of plausibility judgements and validation on multiple text comprehension.

**Conclusion**

Readers form beliefs about unfamiliar but complex scientific debates with little effort. These beliefs are in turn used to evaluate the plausibility or credibility of statements from texts, and quite often readers seem to not make attempts to strategically evaluate and critically reflect on conflicting information. From an individual perspective, this mechanism enables the reader to form an understanding of an unfamiliar topic without much cognitive effort. From an educational perspective, however, this mechanism hampers readers' ability to fully understand and evaluate complex scientific debates, to form well-justified argumentative positions, and to make informed decisions about social, political, or medical topics. Acknowledging the causal role of prior

beliefs, plausibility, and validation in regular comprehension can stimulate further research in educational psychology to shield against belief or plausibility effects and to adapt theories and research on comprehension to the new demands that the World Wide Web imposes on individuals who seek scientific information. Given that scientific controversies are more and more discussed in public on the Web, this issue is of increasing importance.

**References**

3Sat (2018, May, 07th). Starkes Material [Strong material, Video]. Retrieved from https://www.3sat.de/wissen/nano/starkes-material-100.html

Abendroth, J., & Richter, T. (2020a). Text-belief consistency effect in adolescents' comprehension of multiple documents from the Web. *Journal for the Study of Education and Development*.

Abendroth, J., & Richter, T. (2020b). How to understand what you don't believe: Metacognitive training prevents belief-biases in multiple text comprehension. *Manuscript submitted for publication.*

Abendroth, J., & Richter, T. (2020c, November 4th). Mere plausibility enhances comprehension: The role of plausibility in comprehending an unfamiliar scientific debate. Retrieved from osf.io/rvg6b

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/TAC.1974.1100705

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1061–1070. doi:10.1037/0278-7393.19.5.1061

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*, 1037–1049. doi:10.1037/h0077720

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Sigmann, H. (2014). Lme4: linear mixed-effects models using Eigen and S4[Software]. R-package version 1.1- 21. Retrieved in November 2019 from: http://cran.r-project.org/package=lme4

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. doi:10.1037//0022-3514.51.6.1173.

Bartón, B. K. (2020). *Multi-Model inference (MuMin)*. R-package Version 1.43.17. Retrieved June 2020 from https://cran.r-project.org/src/contrib/MuMIn_1.43.17.tar.gz

Beker, K., Jolles, D., Lorch, R. F., & van den Broek, P. (2016). Learning from texts: Activation of information from previous texts during reading. *Reading and Writing*, *29*, 1161-1178. doi: 10.1007/s11145-016-9630-3

Black, A., Freeman, P., & Johnson-Laird, P. N. (1986). Plausibility and the comprehension of text. *British Journal of Psychology, 77*, 51-62.

Blanc, N., Kendeou, P., van den Broek, P., & Brouillet, D. (2008). Updating situation models during reading of news reports: Evidence from empirical data and simulations. *Discourse Processes, 45*, 103–121. doi:10.1080/01638530701792784

Bohn-Gettler, C. M., & McCrudden, M. T. (2018). Effects of task relevance instructions and topic beliefs on reading processes and memory. *Discourse Processes, 55*, 410-431. doi:10.1080/0163853X.2017.1292824

Braasch, J. L. G., & Bråten, I. (2017). The Discrepancy-induced Source Comprehension (D-ISC) Model: Basic assumptions and preliminary evidence. *Educational Psychologist, 52*, 167-181. doi: 10.1080/00461520.2017.1323219

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology, 71*, 499-515. doi:10.1146/annurev-psych-010419-050807

Britt, M. A., Richter, T., & Rouet, J. -F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104–122. doi: 10.1080/00461520.2014.916217

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Connell, L., & Keane, M. T. (2006). A model of plausibility. *Cognitive Science, 30,* 95-120. doi: 10.1207/s15516709cog0000_53

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 19*, 15-18. doi: 10.2307/1268249

Cook, A. E., & Myers, J. L. (2004). Processing discourse roles in scripted narratives: The influences of context and world knowledge, *Journal of Memory and Language, 50*, 268-288. doi: 10.1016/j.jml.2003.11.003

De Pereyra, G., Britt, M. A., Braasch, J., & Rouet, J. -F. (2014). Reader's memory for information sources in simple news stories: Effects of text and task features. *Journal of Cognitive Psychology, 26*, 187-204. doi: 10.1080/20445911.2013.879152.

Ferretti, T. R., Singer, M., & Patterson, C. (2008). Electrophysiological evidence for the time-course of verifying text ideas. *Cognition*, *108*, 881-888. doi: 10.1016/j.cognition.2008.06.002

Gilead, M., Sela, M., & Maril, A. (2019). That's my truth: Evidence for involuntary opinion confirmation. *Social Psychological and Personality Science*, *10*, 393–401. doi: 10.1177/1948550618762300

Green, P. and MacLeod, C. J. (2016), SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493-498. doi:10.1111/2041-210X.12504

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. S*cience*, *304*, 438-441. doi: 10.1126/science.1095455

Hinze, S. R., Slaten, D. G., Horton, W. S., Jenkins, R., & Rapp, D. N. (2014). Pilgrims sailing the Titanic: Plausibility effects on memory for misinformation. *Memory & Cognition*, *42*, 305-324. doi: 10.3758/s13421-013-0359-9.

Isberner, M. -B., & Richter, T. (2013). Can readers ignore implausibility? Evidence for nonstrategic monitoring of event-based plausibility in language comprehension. *Acta Psychologica, 142*, 15-22. doi: 10.1016/j.actpsy.2012.10.003

Isberner, M. -B. & Richter, T. (2014). Comprehension and validation: Separable stages of information processing? A case for epistemic monitoring in language comprehension. In D. N. Rapp & J. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 245-276). Boston, MA: MIT Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1420–1436. doi:10.1037/0278-7393.20.6.1420

Keil, F. C. (2010). The feasibility of folk science. *Cognitive Science, 34*, 826-862. doi: 10.1111/j.1551-6709.2010.01108.x

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163-182. doi: 10.1037/0033-295X.95.2.163

Körner, A., Joffe, S., & Deutsch, R. (2019). When skeptical, stick with the norm: Low dilemma plausibility increases deontological moral judgments. *Journal of Experimental Social Psychology, 84*, 103834. doi:10.1016/j.jesp.2019.103834.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests for random and fixed effects for linear mixed effect models* (lmer objects of lme4 package). R-package version 3.1-0. Retrieved in November 2019 from: http://cran.r-project.org/web/packages/lmerTest/index.html

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction, 27*, 50-62. doi: 10.1016/j.learninstruc.2013.03.001

Lombardi, D., Nussbaum, E. M., & Sinatra, G. M. (2016). Plausibility judgments in conceptual change and epistemic cognition. *Educational Psychologist*, *51*, 35-56. doi: 10.1080/00461520.2015.1113134

Maier, J. & Richter, T. (2013). How nonexperts understand conflicting information on social science issues: The role of perceived plausibility and reading goals. *Journal of Media Psychology, 25*, 14-26. doi: 10.1027/1864-1105/a000078.

Maier, J., & Richter, T. (2016). Effects of text-belief consistency and reading task on the strategic validation of multiple texts. *European Journal of Psychology of Education, 31*, 479–497. doi:10.1007/s10212-015-0270-9

Maier, J., Richter, T., & Britt, M. A. (2018). Cognitive processes underlying the text-belief consistency effect: An eye-movement study. *Applied Cognitive Psychology, 32,* 171-185. doi: 10.1002/acp.3391

Mayer, R. E. (2017) Using multimedia for e-learning. *Journal of Computer Assisted Learning*, 33: 403– 423. doi:10.1111/jcal.12197.

Matsuki, K., Chow, T., Hare, M., Elman, J., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 37*. doi:913-34. 10.1037/a0022964

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *The psychology of learning and motivation, Vol. 51: The psychology of learning and motivation* (pp. 297-384). San Diego, CA, US: Elsevier Academic Press.

Münchow, H., Richter, T., von der Mühlen, S., & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network and relevance for academic success at the university. *British Journal of Educational Psychology, 89*, 502-523. doi:10.1111/bjep.12298

Nieuwenhuis, R., te Grotenhuis, M., & Pelzer, B. (2012). Influence.ME: Tools for detecting influential data in mixed effects models, *The R Journal, 4*, 38-47. doi: 10.32614/RJ-2012-011.

O'Brien, E. J. & Cook, A. E. (2016). Coherence threshold and the continuity of processing: The Ri-Val model of comprehension, *Discourse Processes, 53*, 326-338, doi: 10.1080/0163853X.2015.1123341

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mahwah, NJ: Erlbaum.

Posner, G. J., Strike, K. A., Hewson, P. W. and Gertzog, W. A. (1982), Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*, 211-227. doi:10.1002/sce.3730660207

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, *89*, 250-280. doi: 10.1037/0033-295X.89.3.250

Richter, T. (2011). Cognitive flexibility and epistemic validation in learning from multiple texts. In J. Elen, E. Stahl, R. Bromme, & G. Clarebout (Eds.), *Links between beliefs and cognitive flexibility* (pp. 125-140). Berlin, Germany: Springer.

Richter, T. (2015). Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes*, *52*, 337–355. doi: 10.1080/0163853X.2015.1025665

Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist, 52*, 148-166. doi:10.1080/00461520.2017.1322968

Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology, 96*, 538–598. doi:10.1037/a0014038

Richter, T., & Singer, M. (2017). Discourse updating: Acquiring and revising knowledge through discourse. In D. Rapp, A. Britt, & M. Schober (Eds.), *Handbook of discourse processes (2nd ed.)* (pp. 167-190). New York: Taylor & Francis.

Scharrer, L., Stadtler, M., & Bromme, R. (2014). You'd better ask an expert: Mitigating the comprehensibility effect on laypeople's decisions about science-based knowledge claims, *Applied Cognitive Psychology, 28*, 465–471. doi: 10.1002/acp.3018

Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language, 25*, 279-294. doi: 10.1016/0749-596X(86)90002-1

Schroeder, S., Richter, T., & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language, 59*, 237-259. doi: 10.1016/j.jml.2008.05.001

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464. doi:10.1214/aos/1176344136.

Schwenkenbecher, J. (2019, March 22nd). *Mehr Melken hilft nicht* [More milking does not help]. Retrieved from https://www.sueddeutsche.de/wissen/materialwissenschaft-mehr-melken-hilft-nicht-1.4378492

Singer, M. (2006). Verification of text ideas during reading. *Journal of Memory and Language, 54*, 574–591. doi: 10.1016/j.jml.2005.11.003

Singer, M. (2019). Challenges in processes of validation and comprehension, *Discourse Processes, 56*, 465-483, doi: 10.1080/0163853X.2019.1598167

Thagard P., & Findlay S. (2010). Changing minds about climate change: Belief revision, coherence, and emotion. In E. J. Olsson & S. Enqvist (Eds.), *Belief revision meets philosophy of science, logic, epistemology, and the unity of science* (Vol. 21, pp. 329-345), doi:10.1007/978-90-481-9609-8_14

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2019). Mediation: R package for causal mediation analysis [Software]. R-package version 4.5.0. Retrieved in November 2019, from: http://cran.r-project.org/package=mediation

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, *28*, 247–253. doi:10.1177/0963721419827854

Wertgen, A. G., & Richter, T. (2020). Source information and plausibility interact in the validation of textual information. *Memory and Cognition*. doi: 10.3758/s13421-020-01067-9

Wiley, J. (2005). A fair and balanced look at the news: What affects memory for controversial arguments? *Journal of Memory and Language, 53*, 95-109. doi:10.1016/j.jml.2005.02.001

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301–311. doi:10.1037/0022-0663.91.2.301

Wolfe, M. B., Tanner, S. M., & Taylor, A. (2013). Processing and representation of arguments in on-sided texts about disputed topics. *Discourse Processes, 50*, 457-497. doi: 10.1080/0163853X.2013.828480

Wyer, Jr., R. S., & Radvansky, G. A. (1999). The comprehension and validation of social information. *Psychological Review, 106*, 89–118. doi: 10.1037/0033-295X.106.1.89

van den Broek, P., Beker, K., & Oudega, M. (2015). Inference generation in text comprehension: Automatic and strategic processes in the construction of a mental representation. In E. J. O'Brien, A. E. Cook, & R. F. Lorch (Eds.), *Inferences during reading* (pp. 94–121). Cambridge, UK: Cambridge University Press.

Van der Meer, T., Te Grotenhuis, M., & Pelzer, B. (2010). Influential cases in multilevel modeling. A methodological comment. *American Sociological Review, 75*, 173–178. doi: 10.1177/0003122409359166

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, L.M., & Berthold, K. (2016). Judging the plausibility of argumentative statements in scientific texts: A student-scientist comparison. *Thinking and Reasoning, 22*, 221-249. doi: 10.1080/13546783.2015.1127289

Voss, J. F., Fincher-Kiefer, R., Wiley, J., & Silfies, L. N. (1993). On the processing of arguments. *Argumentation, 7,* 165–181. doi: 10.1007/BF00710663

**Table 1**

*Results from the Pilot-Study about the Perceived Quality of the Videos as Experimental Manipulation*

| | Contra Video Version | | Pro Video Version | |
|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* |
| Understandability [a] | 4.04 | 0.76 | 4.41 | 0.91 |
| Plausibility [a] | 3.44 | 0.88 | 3.91 | 0.89 |
| Quality of Sound [a] | 4.13 | 1.06 | 3.98 | 1.01 |
| Quality of Picture [a] | 3.59 | 0.94 | 3.78 | 0.98 |
| Interestigness [a] | 3.78 | 1.19 | 4.26 | 1.31 |
| Fit Sound-Picture [a] | 4.48 | 1.12 | 3.94 | 1.54 |
| Length [b] | 4.26 | 0.94 | 4.29 | 1.19 |

*Note.* Results of the pilot study with 61 independent university students. Understandability was measured with five items (Cronbach's α = .78), plausibility was measured with six items (Cronbach's α =.88), quality of sound was measured with six items (Cronbach's α = .85), quality of picture was measured with six items (Cronbach's α = .85). Interest, fit between the sound and the pictures, and length were assessed with one item each. [a] Response categories ranging from 1 = *not at all* to 6 = *totally,* [b] Response categories ranging from 1 = *too short* to 6 = *too long.*

**Table 2**

*Text Characteristics of the Two Experimental Texts*

| | Plausibility[a] | Difficulty[a] | Number of Arguments[a] | Clarity of Stance[a] | Interest[a] | Position of the text[b] | Length[c] | Reada-bility[d] |
|---|---|---|---|---|---|---|---|---|
| | M (SEM) | M (SEM) | M (SEM) | M (SEM) | M (SEM) | M (SEM) | | |
| Pro Text | 4.55 (.29) | 5.84 (.19) | 3.83 (.53) | 6.67 (.26) | 5.83 (.21) | 7.00 (.00) | 644 | 37 |
| Contra Text | 4.39 (.29) | 5.5 (.14) | 4.17 (.49) | 5.67 (.45) | 5.58 (.23) | 1.00 (.00) | 685 | 37 |

*Note.* Plausibility = measured with nine items (response categories ranging from 0 = *not at all* to 6 = *totally;* Cronbach's α = .85/.91). Understandability = measured with nine items (response categories ranging from 0 = *not at all* to 6 = *totally;* Cronbach's α = .80/.81). Number of Arguments = number of identified arguments in an open answer question. Clarity of Stance and Interest were assessed with one item each (response categories ranging from 0 = *not at all* to 6 = *totally*).

[a] Results of the pilot-testing with ratings of 12 participants (response categories ranging from 0 = *not at all* to 7 = *totally*), [b] Position of the text (response category ranging from 1= *spider silk should not be used* to 7 = *spider silk should be used*). [c] number of words per text, [d] determined with the German adaption of the Flesch's Reading Ease Index (Amstad, 1978).

**Table 3**

*Examples of Paraphrases, Inferences, and Distracters per Topic Used in the Verification Task (translated into English)*

| Type of test item | Pro text | Contra text |
|---|---|---|
| Original paragraph | […] Due to their unique composition, artificial nerve fibres made from spider silk are very resilient. Computer simulations provide insight into the composition of the spider threads. *These simulations show that spider silk on the one hand consists of tiny soft and unstructured units*. On the other hand, spider silk is also composed of ordered structures. The ordered structures can be imagined like a scaffold with cross and longitudinal beams. They link the unstructured units. […] | […] In addition, milking changes the mechanical properties of the spider silk so that the quality of native spider silk is not achieved. This must be subsequently produced in a complex chemical conversion process by infiltrating the spider silk with metal atoms. *In further processing, the spider threads are then twisted together several times in order to produce a thread with high tensile strength.* The whole milking and conversion process not only takes a very long time in this way. In addition, sufficient quantities of usable spider silk are not available for everyday clinical use. […] |
| Paraphrase | Spider silk is partly composed of small unorganized and elastic particles. | In order to construct a tear-proof thread, the spider threads are repeatedly twisted together in the further course of processing. |
| Inference | The unstructured units of spider silk favor a tension-free connection of severed nerve fibres and make endogenous transplants unnecessary. | If nylon threads are used to connect fine nerves, the risk of reduced conductivity of the nerve tract is reduced. |
| Distracter | A good blood supply to the wound and clean, low-germ wound conditions are the prerequisites for primary wound healing after the connection of nerve fibres. | Decisive for the quality of spider silk is the time when the protein chains are connected with each other. |

*Note.* The paraphrases refer to the italicized sentences in the original paragraphs. In the original texts, the sentences were not italicized. Note that distracters did not directly belong to either the pro or the contra text but are presented in these columns only for illustration.

**Table 4**

*Descriptive Statistics and Intercorrelations.*

| **Measure** | ***M*** | ***SD*** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|---|---|
| 1 Video Version[a] | .04 | 1.01 | 1 | | | | | | |
| 2 Reading Order[a] | -.07 | 1.01 | .00 | 1 | | | | | |
| 3 Task Order[a] | 0.04 | 1.01 | .04 | .00 | 1 | | | | |
| 4 Mean Plausibility Score Pro Text | 0.72 | 0.20 | .27* | -.17 | -.05 | 1 | | | |
| 5 Mean Plausibility Score Con Text | 0.72 | 0.21 | -.22 | .23 | .14 | -.02 | 1 | | |
| 6 Mean Comprehension Score Pro Text | 0.77 | 0.21 | .02 | .21 | -.27* | .65*** | .08 | 1 | |
| 7 Mean Comprehension Score Con Text | 0.76 | 0.19 | -.27* | .05 | -.17 | .25 | .51*** | .46** | 1 |

*Note*. $N = 54$.

[a] contrast coded; Video Version (-1 = contra video version, 1 = pro video version), Reading Order (-1 = contra text first, 1 = pro text first), Task Order (validation task first = -1, verification task first = 1). Mean Comprehension Score: Mean accuracy responses averaged across paraphrases and inferences in the verification task. Mean Plausibility Score: Mean plausibility judgments averaged across paraphrases and inferences in the validation task.

* $p \leq 0.05$, ** $p < .01$, *** $p < .001$ (two-tailed).

**Table 5**

*Fixed Effects, Variance Components and Fit Statistics of the Three GLMMs for Steps 1 to 3.*

| | *Step 1*: Effects of text type and video version on comprehension | *Step 2*: Effects of text type and video version on plausibility judgments | *Step 3*: Effects of text type, video version and perceived plausibility on comprehension |
|---|---|---|---|
| | | Fixed Effects | |
| Parameter | β (*SE*) | β (*SE*) | β (*SE*) |
| Intercept | 1.52 (0.18)*** | 1.14 (0.16)*** | 1.23 (0.17)*** |
| Text Type[a] | 0.04 (0.13) | -0.00 (0.13) | 0.05 (0.13) |
| Video Version[a] | -0.15 (0.14) | 0.04 (0.11) | -0.17 (0.13) |
| Participants' Responses to Distracters[b] | -0.07 (0.14) | -0.02 (0.11) | -0.05 (0.13) |
| Reading Order[a] | 0.21 (0.14) | 0.06 (0.11) | 0.22 (0.13) |
| Task Order[a] | -0.34 (0.14)* | 0.05 (0.11) | -0.40 (0.13)** |
| Text Type × Video Version | 0.19 (0.07)** | 0.30 (0.06)*** | 0.07 (0.07) |

| | | | |
|---|---|---|---|
| Video Version × Reading Order | 0.29 (0.14) | 0.13 (0.11) | 0.15 (0.13) |
| Video Version × Task Order | 0.22 (0.14) | 0.11 (0.11) | 0.20 (0.13) |
| Reading Order × Task Order | -0.04 (0.14) | -0.01 (0.11) | -0.04 (0.13) |
| Video Version × Reading Order × Task Order | 0.03 (0.14) | -0.09 (0.11) | 0.07 (0.13) |
| Perceived Plausibility[a] | | | 0.95 (0.08)*** |

| | Variance Components | | |
|---|---|---|---|
| Parameter | Variance (*SD*) | Variance (*SD*) | Variance (*SD*) |
| Subjects | 0.70 (0.84) | 0.41 (0.64) | 0.56 (0.75) |
| Test Items | 0.40 (0.63) | 0.38 (0.62) | 0.32 (0.56) |

*Note.* $N = 1541$ (54 participants x 30 items minus outliers)

[a] contrast coded; [b] grand-mean centered. Text Type (-1 = contra text, 1 = pro text), Video Version (-1 = contra video version, 1 = pro video version), Reading Order (-1 = contra text first, 1 = pro text first), Task Order (validation task first = -1, verification task first = 1). Perceived Plausibility (Predictor in Step 3: -1 = implausible, 1 = plausible). Comprehension: Accuracy responses to paraphrases and inferences in the verification task (0 = incorrect answer, 1 = correct answer). Plausibility (Dependent variable in Step 2): Plausibility judgments to paraphrases and inferences in the validation task (0 = implausible, 1 = plausible).

* $p < 0.05$, ** $p < .01$, *** $p < .001$ (two-tailed).

**Table 6**

*Comparison of Structural Models for Comprehension (Verification Task) and Plausibility Judgements (Validation Task).*

| | | Parameter | | | | | Model Comparison | | |
|---|---|---|---|---|---|---|---|---|---|
| Outcome | Model | *Log-Likelihood* | *df* | *AIC* | *BIC* | *Deviance* | | Δ *df* | Δχ 2 | *p* |
| Comprehension | | | | | | | | | | |
| | Model 1a | -756.9 | 3 | 1519.8 | 1535.8 | 1513.8 | | | | |
| | Model 2a | -735.5 | 7 | 1521.0 | 1558.4 | 1507.0 | Model 2a vs. Model 1a | 4 | 6.8 | .15 |
| | Model 3a | -746.9 | 13 | 1519.7 | 1589.1 | 1493.7 | Model 3a vs. Model 1a | 10 | 20.1 | .03* |
| | | | | | | | Model 3a vs. Model 2a | 6 | 13.3 | .04* |
| | Model 4a | -667.9 | 14 | 1363.9 | 1438.7 | 1335.9 | Model 4a vs. Model 1a | 11 | 177.9 | <.001*** |
| | | | | | | | Model 4a vs. Model 2a | 7 | 171.7 | <.001*** |
| | | | | | | | Model 4a vs. Model 3a | 1 | 157.8 | <.001*** |
| Plausibility Judgements | | | | | | | | | | |
| | Model 1b | -864.2 | 3 | 1734.5 | 1750.5 | 1728.5 | | | | |
| | Model 2b | -864.0 | 7 | 1741.9 | 1779.3 | 1727.9 | Model 1b vs. Model 2b | 4 | 0.6 | .97 |
| | Model 3b | -850.3 | 13 | 1726.7 | 1796.1 | 1700.7 | Model 3b vs. Model 1b | 10 | 27.8 | .002** |
| | | | | | | | Model 3b vs. Model 2b | 6 | 27.3 | <.001*** |
| | Model 4b | -768.7 | 14 | 1565.4 | 1640.2 | 1537.4 | Model 4b vs. Model 1b | 11 | 191.1 | <.001*** |
| | | | | | | | Model 4b vs. Model 2b | 7 | 190.5 | <.001*** |

| | | | |
|---|---|---|---|
| Model 4b vs. Model 3b | 1 | 163.2 | <.001*** |

*N* = 1541 (54 participants x 30 items minus outliers)

*Note*. The null Models 1a and 1b contained only the random effects for participant and test item. In the baseline Models 2a and 2b, the control factors reading order (-1 = contra text first, 1 = pro text first), task order (validation task first = -1, verification task first = 1), the interaction of reading order and task order as well as participants' responses to distracters (grand-mean centered) were entered as fixed effects. In the unmediated Models 3a and 3b, text type (-1 = contra text, 1 = pro text), video version (-1 = contra video version, 1 = pro video version), their interaction and the interactions of the between-subject factors were added as fixed effects. In the mediated Models 4a and 4b, the additional fixed effect of the proposed mediator was included (i.e., fixed effect of plausibility (-1 = implausible, 1 = plausible) in Model 4a; fixed effect of comprehension accuracy (-1 = incorrect answer, 1 = correct answer) in Model 4b).

* $p < 0.05$, ** $p < .01$, *** $p < .001$ (two-tailed).

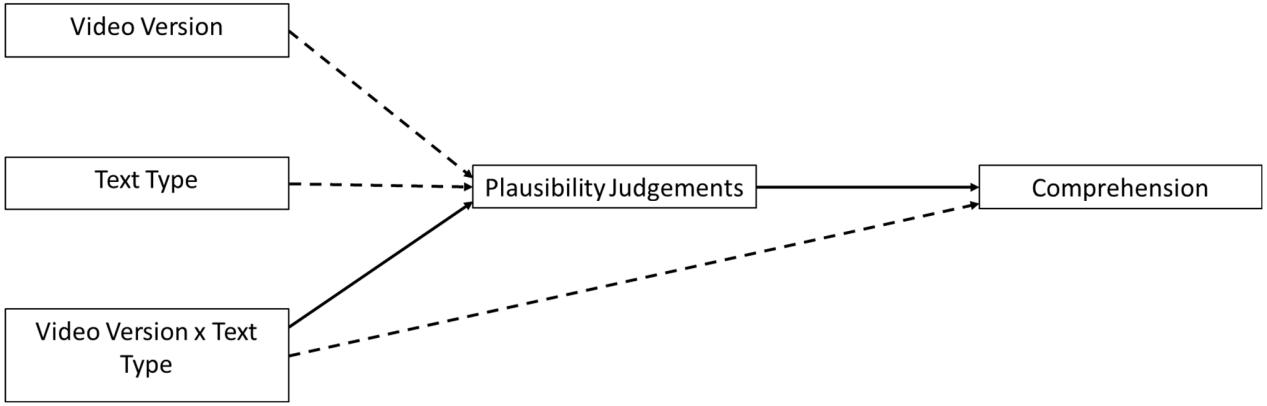


*Figure 1.* Mediation model of video version, text type and plausibility on comprehension. Solid lines indicate the relationship in question of the mediation model, that is, the hypothesized indirect relationship of the interaction of video version and text type on comprehension mediated by plausibility. Dashed lines indicate direct effects of the independent variables on the mediator plausibility and on the outcome comprehension.

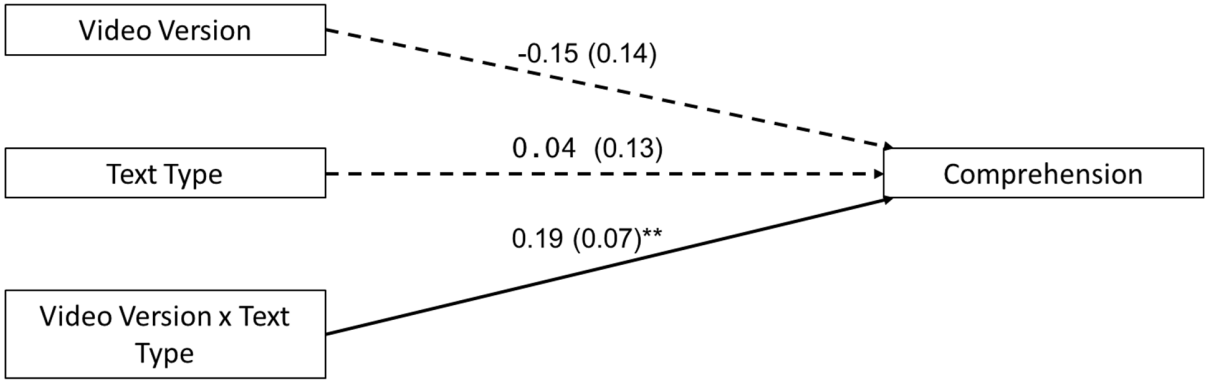


*Figure 2.* Results of the unmediated effects of text type (-1 = contra text, 1 = pro text) and video version (-1 = contra video version, 1 = pro video version) on comprehension (percentage of accurate responses to paraphrases and inferences in the verification task). Parameter estimates are based on the generalized linear mixed model (GLMM) analysis with logit link function from Step 1 and standard errors are provided in parentheses. Solid lines indicate the relationship in question for Step 1, that is, the hypothesized relationship of the interaction of video version and text type on comprehension. Dashed lines indicate the direct effects of the independent variables on comprehension. For clarity of presentation, effects of reading order and task order are not visualized.

** $p < .01$

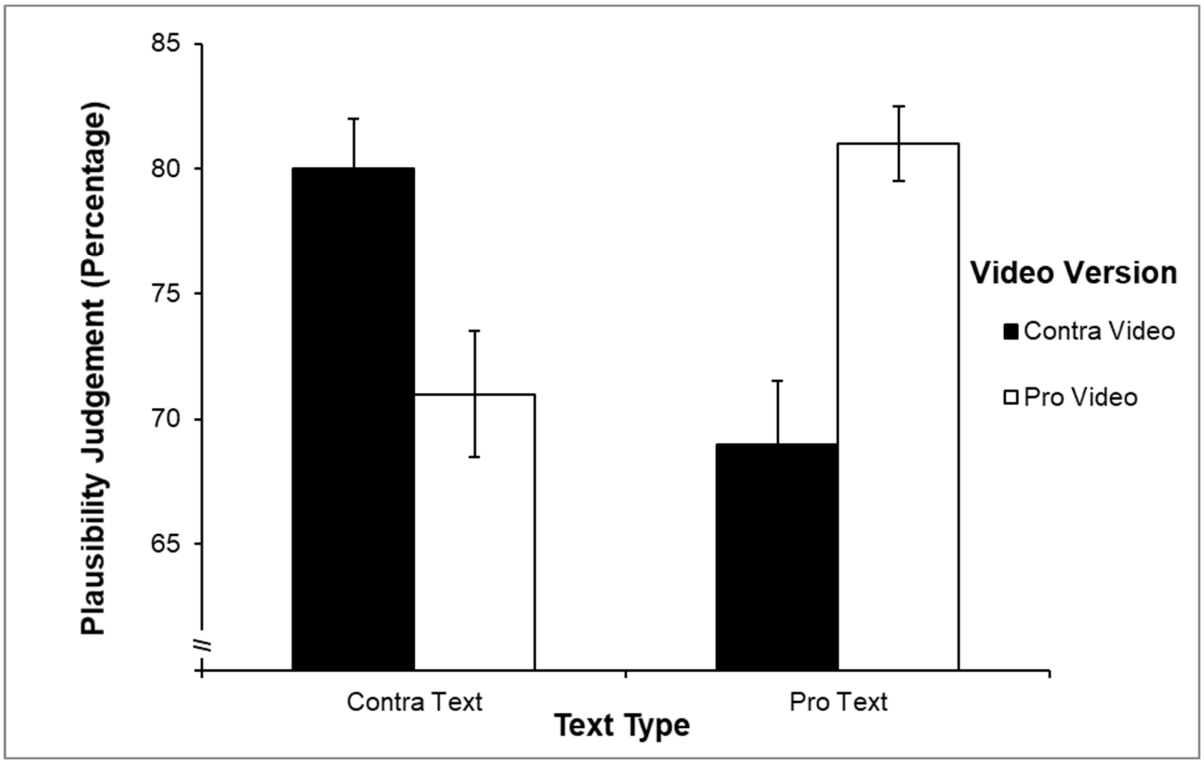


*Figure 3*. Interaction effects of text type (contra vs. pro) and video version (contra vs. pro) on plausibility judgements (percentage of plausibility judgments to paraphrases and inferences in the validation task). Probabilities were back-transformed from the generalized linear mixed model (GLMM) analysis with logit link function from Step 2 and are provided with estimated standard errors.

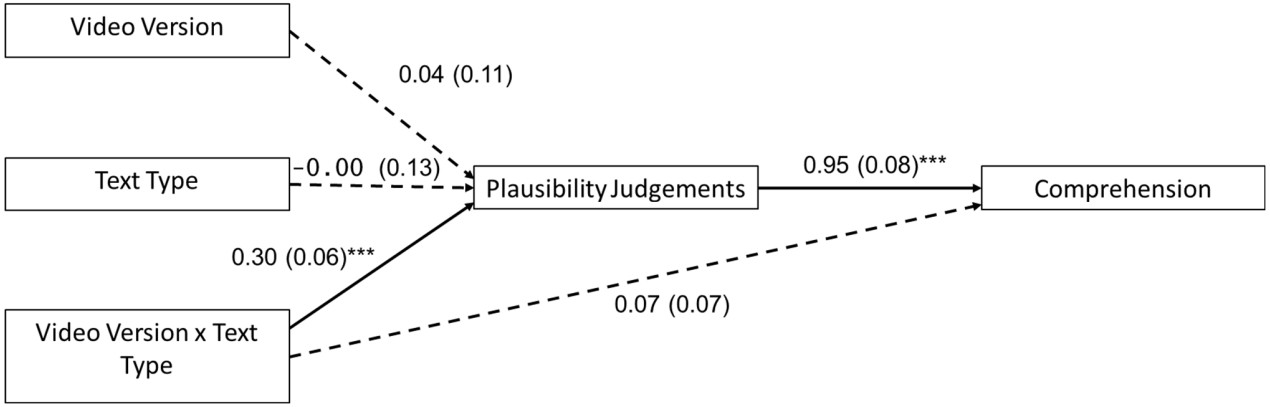


*Figure 4.* Results of the mediation model for the effect of text type (-1 = contra text, 1 = pro text) and video version (-1 = contra video version, 1 = pro video version) with perceived plausibility (-1 = implausible, 1 = plausible) as mediator on comprehension (percentage of accurate responses to paraphrases and inferences in the verification task). Parameter estimates are based on the generalized linear mixed model (GLMM) analysis with logit link function from Steps 2 and 3. Standard errors are provided in parentheses. Solid lines indicate significant relationships and dashed lines paths that were not significant in the GLMM analyses. For clarity of presentation, effects of reading order and task order are not visualized. *** $p < .001$