Accepted for publication in Journal of Experimental Psychology: General

Correction of Evident Falsehood Requires Explicit Negation

Rebecca Weil, Yaacov Schul, & Ruth Mayo

The Hebrew University of Jerusalem

Author Note

Rebecca Weil, The Martin Buber Society of Fellows, The Hebrew University of Jerusalem. Yaacov Schul and Ruth Mayo, Department of Psychology, The Hebrew University of Jerusalem.

Rebecca Weil is now at the Department of Psychology, Faculty of Health Sciences, University of Hull, HU6 7RX, UK.

This work was supported by the Israel Science Foundation (ISF 594/12) and by a fellowship of The Martin Buber Society of Fellows, The Hebrew University of Jerusalem, sponsored by the Federal Ministry of Education and Research, Germany. The authors thank Noam Siegelman and Eric D. Gardiner for their help conducting the linear mixed-model analyses. The ideas and data presented in this article were previously presented at the Language and Social Cognition Conference (2015), Bern, Switzerland; the Third Conference on Cognition Research (2016), Israeli Society for Cognitive Psychology, Akko, Israel; the Cologne Social Cognition Meeting (2018), Germany; and at research talks at the University of Hull, UK.

Correspondence concerning this article should be addressed to Rebecca Weil, Department of Psychology, Faculty of Health Sciences, University of Hull, HU6 7RX, UK. Contact: r.weil@hull.ac.uk

Wordcount: [15200]

Abstract

The danger of receiving false information is omnipresent, and people might be highly vigilant against being influenced by falsehoods. Yet, as research on misinformation reveals, people are often biased by false information, even when they know the valid alternative. The question is why? The current research explores the relative encoding strength of two opposing alternatives involved in the correction of falsehood: the false concept and the valid concept. These encoding strengths may be critical for what people remember and how they act upon receiving false information. We compared two triggers for the correction of falsehood—a sentence consisting of clearly false information (e.g., "honey is made by butterflies") and a sentence consisting of an explicit negation of this information (e.g., "honey is not made by butterflies"). The general pattern of results from five experiments demonstrates that the valid concept (e.g., "bees") exhibits a weaker presence in memory than the false concept (e.g., "butterflies") following the comprehension of evidently false information as compared to its explicit negation. Thus, the current research provides an answer to the riddle of the persistence of false information: False information is less likely to be mentally corrected if it is not explicitly negated. Even when people detect that a sentence is false, they tend to focus on the false concept rather than on the valid concept. These findings shed new light on extant research and offer fresh insights about the processing of false information and related phenomena such as the reliance on misinformation.

Keywords: negation, validation, memory, false information, misinformation

"The best way to get the right answer on the Internet is not to ask a question, it's to post the wrong answer." The claim, denoted as Cunningham's Law (McGeady, 2010; but see Cunningham, 2015), offers an optimistic view of the crisis of deceptive communication, suggesting that it is actually easier to arrive at a correct answer in response to a false sentence than in response to a question. In an age of online misinformation, in which fake news might spread faster and further than true information (Lazer et al., 2018), Cunningham's Law appears outdated and the belief in its validity potentially dangerous, allowing for a dissemination of falsehoods. The zeitgeist is pessimistic, highlighting the fact that people do not cope well with falsehoods and often fail to correct false information (e.g., Fazio, Barber, Rajaram, Ornstein, & Marsh, 2013; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Pantazi, Kissine, & Klein, 2018). The present research investigates the correction of false information by comparing two potential triggers for the correct answer—a clearly false sentence versus an explicit negation of falsehood.

To illustrate, assume you are confronted with the sentence that "honey is made by butterflies." Presumably, you noticed immediately with minimal effort that this sentence is false and rejected it. Do you just reject the false sentence or do you mentally correct the information by thinking of the valid concept "bees," instead of the false concept "butterflies"? The present research tests this question in comparison to the outcomes of an explicit negation of false information as, for example, when you are told that "honey is *not* made by butterflies." Here too, we ask whether people go beyond the negation and think of the valid concept "bees." The two central questions of the present research are whether clearly false sentences and the negation of clearly false sentences allow, to the same degree, for the encoding of valid concepts and whether the two cases differ with respect to the encoding of false concepts.

The encoding strength of the valid concept (e.g., bees) relative to the false concept (e.g., butterflies) that is present in the false communication, may be critical for how people act

upon receiving false information and what they remember. The theoretical analysis of this issue entails integration of research on negation and validation into a larger theoretical framework. In doing so, we advocate a theoretical and methodological approach, attempting to highlight the challenges involved in handling false information.

A Comparison of Falsehood and Negation

Validation of False Information

Validation is one of the cornerstones of meaning negotiation that occurs when recipients deal with incoming information. It is a by-product of the process of comprehension, generated upon receiving information (e.g., Richter 2015). In the case of obviously false sentences (e.g., "Honey is made by butterflies"), one recognizes that the sentence is false because it contradicts one's existing knowledge (e.g., "Honey is made by bees"; see Kintsch & Van Dijk, 1978; Myers & O'Brien, 1998; Richter, Schroeder, & Wöhrmann, 2009; Singer, 2013; Van Dijk & Kintsch, 1983). Because the danger of receiving false information is increasingly present in a culture that promulgates fake news (e.g., Hunt, 2016), people might be highly vigilant in detecting falsehoods (cf. Conroy, Rubin, & Chen, 2015). Indeed, the research on validation shows that people are proficient at detecting falsehood (Cook & O'Brien, 2014; Isberner & Richter, 2013, 2014; Richter et al., 2009).

Yet, surprisingly, despite the findings attesting to the proficiency of validation processes, accumulating evidence indicates that people do in fact adopt false information even when they have existing knowledge that should have allowed them to reject it (for an overview, see Rapp & Braasch, 2014; see also earlier research on belief perseverance reviewed in Schul & Burnstein, 1998). This is commonly referred to as misinformation effect (e.g., Fazio et al., 2013; Singer 2019). At the most general level, the misinformation effect occurs when people are influenced by false information in spite knowing that it is false.

However, the general term misinformation effect may mask the complexity of the phenomenon, and in fact, there might be important variations in the conditions that bring about susceptibility to false information (e.g., reliance on misinformation despite contradicting prior knowledge, continued influence of information that has been labeled as false, memory distortion of post-event misinformation; see also Ecker, Lewandowsky, Cheung, & Maybery, 2015). In the present manuscript we use the term reliance on misinformation to refer to biases caused by false information that contradict peoples' prior knowledge. A typical finding is that even people who reveal factual knowledge in an early test (e.g., "The largest ocean is the Pacific") err in answering the question "What is the largest ocean?" after being exposed to misinformation (e.g., "The largest ocean is the Atlantic"). The reliance on misinformation highlights the necessity of understanding the conditions under which validation takes place (for a meta-analysis on the efficacy of debunking misinformation, see Chan, Jones, Hall Jamieson, & Albarracín, 2017; see also Lewandowsky, Ecker, & Cook, 2017) and how validation influences the encoding of the valid concept relative to the encoding of the presented false concept (Rapp, 2016; Singer & Doering, 2014).

Hinze, Slaten, Horton, Jenkins, and Rapp (2014) proposed that the plausibility of the false concept in the given context influences the encoding when false sentences are processed. These authors reported that false yet plausible concepts (e.g., "The Pilgrims' ship was the *Godspeed*") were encoded more strongly than false and implausible concepts (e.g., "The Pilgrims' ship was the *Titanic*"). The plausibility of the context is important as well. Rapp, Hinze, Slaten, and Horton (2014) showed that a realistic context led to a stronger encoding of false concepts as compared to an unrealistic context, suggesting that contextual information influences whether false or valid concepts should be prioritized. Indeed, Rapp (2008) demonstrated that even when participants were explicitly instructed to generate valid information prior to receiving false information, such instructions did not overcome the influence of a suspenseful context and did not substantially aid the validation process.

Another factor that might affect the encoding strength of false relative to valid concepts is source credibility (Rapp, 2016). The assumption here is that people rely more on trustworthy sources and discount information from untrustworthy sources (see Chaiken & Maheswaran, 1994; Hovland & Weiss, 1951, for early conceptualizations; Brinol & Petty, 2009; Smith, De Houwer, & Nosek, 2013, for more recent discussions). Thus, comparing the valid and false concepts that are triggered by the same sentence, it could be assumed that valid concepts are encoded more strongly when false information is provided by untrustworthy sources, and false concepts are encoded more strongly when the false information is provided by trustworthy sources. Yet, Sparks and Rapp (2011) demonstrated that readers consider the credibility of a source only after they have comprehended information and evaluated its consistency with the active memory contents. Accordingly, source credibility might not influence the initial encoding of the information, but rather, encoding might be modified after validation is completed (see also Nadarevic & Erdfelder, 2013). Finally, encoding of the valid concept when encountering false information appears to be particularly likely when recipients are primed to think of alternatives, as, for example, when they generate counter-information (e.g., Xu & Wyer, 2012) or they are in the mindset of distrust (Mayo, 2015; Schul, Mayo, & Burnstein, 2004).

Negation of False Information

In natural conversations, negations are as prevalent as words connoting positive emotions, twice as frequent as words connoting negative emotions, and almost three times more prevalent than words denoting causality (Pennebaker, Mehl, & Niederhoffer, 2003). The dominant view is that processing negations is a complex task that demands cognitive resources (e.g., Grant, Malaviya, & Sternthal, 2004) and often leads to memory failure (Carpenter & Just, 1975; Clark & Chase, 1972; Horn, 1989; Johnson-Laird & Savary, 1999; Just & Carpenter, 1976; Lea & Mulligan, 2002; see Mayo, 2015, for exceptions). Moreover, the use of negations entails higher syntactic and semantic complexity and hence involves the

risk of misunderstanding compared to affirmative phrasings with the same meaning (Colston, 1999; Fiedler, Walther, Armbruster, Fay, & Naumann, 1996; Mayo, Schul, & Burnstein, 2004). So why do communicators choose to use a negation instead of an affirmative phrasing?

Negations are attributed specific functions, such as mitigation and politeness (Brown & Levinson, 1987; Fraenkel & Schul, 2008), contradiction of expectations or beliefs, as done by denial or rejection (Clark & Clark, 1977; Givón, 1993; Horn, 1989), and the conveyance of understatement or irony (e.g., Giora, Balaban, Fein, & Alkabetz, 2005). Although affirmation and negation can often be used interchangeably (Giora, 2006; but see Beltrán, Orenes, & Santamaría, 2008), people sometimes prefer negation to ensure unambiguous communication, especially in the case of the rejection of a particular wrong message (Givón, 1978). To illustrate, although the vast majority of people would agree that honey is made by bees, the use of negation in the sentence "Honey is *not* made by butterflies" implies that someone assumed or stated that honey *is* made by butterflies. Accordingly, one of the functions of negation is highlighting falsehood. But does the negation induce mental correction? Do recipients encode the valid concept relative to the false concept more strongly when they are confronted with an explicit negation of falsehood?

To the extent that negation lowers the activation level of a negated concept (de Vega et al., 2016; Kaup, 2001; Kaup, Zwaan, & Lüdtke, 2007; MacDonald & Just, 1989; see also Mayo, Schul, & Rosenthal, 2014), negation could be considered an effective counter-arguing linguistic tool in that it reduces thinking about the falsehood. However, often the opposite is found, and in actuality negations can lead to stronger encoding of false concepts compared to valid ones (e.g., Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008).

Research shows that when valid concepts are not accessible or not available (see Mayo et al., 2004; Schul & Mayo, 2014), negation might backfire, leading to an encoding of false concepts (e.g., Adriaanse, van Oosten, de Ridder, de Wit, & Evers, 2011; Deutsch,

Gawronski, & Strack, 2006; Deutsch, Kordts-Freudinger, Gawronski, & Strack, 2009; Kaup, Lüdtke, & Zwaan, 2006; Kaup & Zwaan, 2003). However, when a false concept with a clear alternative concept is negated (e.g., Anderson, Huette, Matlock, & Spivey, 2010; Orenes, Beltrán, & Santamaría, 2014), negation might be encoded in line with the intended meaning of the negation (i.e., valid concept).

In particular, the balance of encoding strength of valid relative to false concepts following negated falsehoods seems to be influenced by the pragmatic informativeness of the sentence and recipients' expectations while they process it (Dale & Duran, 2011; Nieuwland & Kuperberg, 2008; Nordmeyer & Frank, 2014). For example, embedding a negated sentence like "cars have no wings" in a context that makes the negation informative and expected (i.e., "Flying cars?! But cars have no wings") might lead to a stronger encoding of valid concepts as compared to false concepts. Tian, Ferguson, and Breheny (2016) proposed that the pragmatic question that a recipient tries to answer when confronted with negation determines whether the recipient favors the false concept over its valid alternative. Finally, the balance between encoding strength of the valid and false concept may depend on personal affordances. Haran, Mor, and Mayo (2011) showed that chronically depressed individuals encoded negated affectively negative concepts more strongly than negated affectively positive ones, suggesting that accessibility of valid concepts might be determined by personal affordances. Importantly, as the above discussion suggests, the availability of a valid concept is a necessary condition for it to be encoded; however, this availability might not be sufficient.

The Encoding of Affirmative False versus Negated True Information

Several studies compared the processing and mental representation of affirmative and negated true and false information within one paradigm (e.g., Clark & Chase, 1972; Hasson & Glucksberg, 2006; Kaup et al., 2006; Nieuwland & Kuperberg, 2008). Consider first the comparison of affirmative true and affirmative false sentences. It has been found that the

former has a processing advantage. To illustrate, assume participants are presented with sentences (e.g., “The star is above the plus sign”) and corresponding pictures, and they have to indicate whether the sentence correctly describes the picture. When the affirmative sentences match the content of the picture, the task is easier than when they mismatch. This reflects the relative ease with which one processes true information as compared to false information. Now consider the comparison of negated true and negated false sentences. In this case the latter has a processing advantage (but see Hasson & Glucksberg, 2006; Nieuwland & Kuperberg, 2008). Thus, for example, when the picture shows a star above a plus sign, it is easier for participants to detect falsehood for the negated false sentence “The star is not above the plus sign” than to verify the negated true sentence “The plus sign is not above the star.” Such findings are typically explained by the suggestion that people initially process the core of a sentence (e.g., star above plus sign), which is then negated. When the core is congruent with the picture it is easy to detect that a negated sentence is false. When the core is incongruent with the picture it becomes more difficult to see that a negated sentence is true (but see Nieuwland & Kuperberg, 2008). However, after a short delay (i.e., 1000–1500 ms; Hasson & Glucksberg, 2006; Kaup et al., 2006), the processing of negated true information does not differ from affirmative true information, indicating that the intended meaning of negation has been processed.

Thus, it is likely that when people comprehend sentences involving a negation of false information, the false concept as well as the corresponding valid concept receive some level of activation. This might also be the case for affirmative sentences that convey false information (e.g., “honey is made by butterflies”). That is, when a false sentence is comprehended, the presented false concept (butterflies) is initially activated and subsequently integrated with other information in memory (Cook & O'Brien, 2014). This process might also involve activation of the valid concept (bees), and hence, it allows for a detection of falsehood by recognizing the mismatch between the false and the valid concept. Accordingly,

the question of interest is whether comprehending a false sentence and an explicit negation of falsehood differs in its influence on the relative encoding strengths of false and valid concepts and thus affects how this information is remembered.

The Present Study

Our research investigates the encoding of valid (e.g., "bees") and false (e.g., "butterflies") concepts following false sentences phrased as affirmations (e.g., "honey is made by butterflies") and their explicit negations (e.g., "honey is not made by butterflies"). For the sake of brevity, we sometimes refer to these sentences simply as false sentences and negated sentences. We use memory of valid and false concepts as a proxy for the encoding strength during comprehension. The influence of negated versus false sentences is assessed relative to a benchmark of baseline sentences (e.g., "Honey is made out of nectar") that include information about the focal object (e.g., honey) without referring to the target concept (e.g., butterflies or bees). Therefore, the baseline sentences allow us to assess the general tendency to think about the target concepts given the focal object in a sentence.

In none of the experiments do we ask participants to make any explicit truth evaluation; however, the experimental encoding tasks do differ regarding the relevance of truth-values. In Experiments 1a and 1b we investigate whether the encoding strength of valid and false concepts differs after false and negated sentences were processed within a grammatical task in which the truth-value of sentences is irrelevant for the task. In Experiments 2a and 2b we introduce an impression-formation task in which the truth-value of sentences has implications for the evaluation of sentences and sources. Experiment 3 separates the influence of the sources from the influence of sentences. Replicating the effect found in the earlier experiments, Experiment 3 demonstrates the importance of the evaluative component of sentences, independent of the sources.

We collected the data of each experiment presented in the current manuscript in one shot without prior statistical analyses. Sample sizes were determined beforehand, and sensitivity analyses (GPower 3.1.9.2) assuming a power of (1-β) = .80 revealed that the experiments were sufficiently sensitive to detect effect sizes of $\eta_p^2 > .01$, for the main statistical effects of interest. We report all data exclusions, all manipulations, and all measures. Materials and data are available at https://osf.io/4xevs/. Experiments 1a, 1b, 2a, and 2b were approved by the ethics committee of the Hebrew University of Jerusalem, and Experiment 3 was approved by the ethics committee of the University of Hull.

Experiment 1a

Experiment 1a employed a task that required participants to process sentences to the degree that they could decide whether the sentences are presented in a correct grammatical order as opposed to being scrambled (i.e., a grammar task). We assumed that detecting a correct or incorrect grammatical order would require participants to read the full sentences but has little to do with the encoding of either false or valid concepts. The question of interest is whether upon processing negated or false sentences, would participants access and encode the corresponding valid concept? Importantly, we selected sentences for which the vast majority of our respondents' population knows the correct answer (see Appendix for a list of all experimental sentences). Thus, we assumed that for all our sentences, the valid concept is potentially accessible.

Experiment 1a tested whether the processing of negated and false sentences within a grammar task would lead to an encoding of valid concepts. In line with previous research, this should be the case for negated sentences (see Mayo et al., 2004; Nieuwland & Kuperberg, 2008) as well as false sentences (see Isberner & Richter, 2013, 2014; Richter et al., 2009). Thus, Experiment 1a investigates the hypothesis that processing of negated and false

sentences leads to a stronger encoding of the valid concept compared to the baseline sentences.

Participants and Design

Because this is a new paradigm, we determined the sample size beforehand, with a total requirement of 250 participants. We recruited 250 participants (162 female, 85 male, 1 other, 2 not reported; M_{age} = 36.85 years) via Amazon's Mechanical Turk (MTurk; e.g., Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013). Participants could sign up for the experiment only if they resided in the United States. We required that they had previously completed at least 100 tests via MTurk and held a record of supplying acceptable data at least 95% of the time. They received $1 for their participation. In a post-experimental demographic questionnaire, we asked about the participants' native language, whether they were interrupted during the experiment or were in the presence of others while performing the task, and whether they had any educated guess concerning the purpose of the experiment.[1] The experiment had a single-factor design with a three-level within-participant variation on the Sentence Type factor (affirmative false vs. negated true vs. baseline true).

Procedure

The experiment consisted of three consecutive phases: a grammar task, a filler task, and a memory-test phase. Stimulus presentation and response collection were controlled by Inquisit 4.0.6.0. In each trial of the grammar task, participants were presented with a sentence that appeared centered in the upper part of the screen. We instructed them to indicate whether the word order was jumbled, by pressing "q," or correctly arranged, by pressing "p." The sentence stayed on the screen until participants indicated their answer (see Figure 1).

[1] For exploratory reasons, the post-questionnaire of Experiment 1a contained a six-item dispositional trust questionnaire (Yamagishi & Yamagishi, 1994).

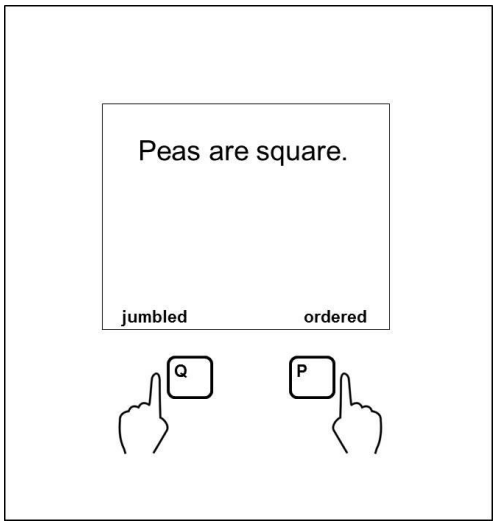

Figure 1. Example of trial of the grammar task in Experiments 1a and 1b.

After the grammar task, participants were presented with a two-minute distraction task, which was modeled after the "Where's Waldo?" series. Upon being shown a complex cartoon-like picture depicting a variety of people doing many different things, participants were asked to find a particular person (i.e., Waldo) in that picture.

In the third and last phase of the experiment, participants took a surprise memory test for words presented in the grammar task. In each memory trial, participants were shown a probe word that appeared centered in the upper part of the screen. We instructed them to indicate on an 8-point scale (1 = *new*, 8 = *old*), which was superimposed below the probe word, whether they thought a probe appeared previously in the grammar task and was hence an "old" word, or whether they thought the probe did not appear before and was hence a "new" word. We explicitly told participants that their response should reflect their old/new categorization as well as their confidence in this categorization. Specifically, they were told: "If you think that the word is new and you are very sure about it, you should press 1; if you are somewhat sure, press 2 or 3; and if you are unsure, press 4. If you think that the word is old and you are very sure about it, you should press 8; if you are somewhat sure, press 6 or 7; and if you are unsure, press 5." The instructions about how to use the scale were designed to

make participants aware of the difference between thinking that a probe is old but being unsure (5) versus thinking that a probe is new but being unsure (4). To emphasize this distinction, labels were placed above the scale. The label "new" appeared on the left side of the scale (above the response options 1 and 2), and the label "old" appeared on the right side of the scale (above the response options 7 and 8). The center of the scale was labeled with "unsure" (placed centered between the response options 4 and 5), optically dividing the scale into response options associated with "new" responses (left side of the scale) and response options associated with "old" responses (right side of the scale) of different levels of confidence.

Materials

Stimuli (grammar task). Each participant saw 120 different sentences. Sixty of the 120 sentences were shown in the experimental trials; the other 60 were fillers. The filler sentences were identical for all participants and were presented in a jumbled order (e.g., "Daffodils flowers are"). The experimental sentences were in Standard English. The experimental items were constructed in the following way: One member of the research team constructed a pool of clearly true sentences. Two other members of the research team selected from this pool the 60 most evidently true sentences, involving simple declarative facts (e.g., "Honey is made by bees"). These sentences were never shown in the experiment. Based on each true sentence, we created three variations: affirmative false sentences (e.g., "Honey is made by butterflies"), negated true sentences (e.g., "Honey is not made by butterflies"), and true baseline sentences (e.g., "Honey is made out of nectar"). Each participant saw only one of the versions, and we counterbalanced among participants the appearance of all three versions. Thus, participants saw 60 experimental sentences: 20 false, 20 negated, and 20 baseline (see Appendix for sentences and counterbalancing conditions). The 60 filler trials aimed to eliminate the association between an affirmative/negated phrasing and the truth-value of the sentence. To this end, the fillers consisted of 20 affirmative true sentences (e.g.,

"Daffodils flowers are"), 20 negated false sentences (e.g., "Barcelona in Spain is not"), and 20 affirmative false sentences (e.g., "Strawberries thorns have"), all in a jumbled order, as mentioned above. Thus, combining the experimental and filler trials, each participant saw 40 affirmative false, 40 affirmative true, 20 negated false, and 20 negated true sentences. Note that unlike the experimental sentences, each of which had three variations (false, negation, and baseline), the filler sentences had only one version each, and all participants saw exactly the same fillers.

Stimuli (memory test). Participants viewed 120 different single-word memory probes, one for each sentence they saw during the grammar task. The memory probes for all the experimental items were "new." That is, they were not presented in any of the sentences. Rather, they consisted of the word that appeared at the end of the original true sentence, which served as the basis for creating the three versions of the experimental items. To illustrate, assume the original sentence (not shown to participants) was "Honey is made by bees." During the grammar task (i.e., in phase 1), each participant saw a false sentence (e.g., "Honey is made by butterflies") or a negated sentence (e.g., "Honey is not made by butterflies") or a baseline sentence (e.g., "Honey is made out of nectar"). All participants were shown the probe word "bee." We tested whether recognition of the valid concept (e.g., "bees") differs as a function of the type of sentence. The memory test rests on the assumption that stronger encoding causes a greater bias in memory, leading participants to indicate that the probe word was presented during the grammar task.

In addition to the 60 experimental memory probes, we presented 40 memory probes consisting of words randomly taken from the filler trials, and 20 new words that had no direct relation to the sentences shown during the grammar task. Accordingly, the 40 words taken from the filler trials are considered "old" and the 20 words with no direct relation to the sentences "new." All in all, participants saw 80 words that were not presented to them before (60 related to the experimental sentences and 20 completely new) and 40 old words. The 120

words were presented in a randomized order. Each participant started the memory task with the same 10 additional filler probes (6 old words, 4 new words), resulting in a presentation of 130 words in total. We did not include the 10 additional filler words in the analysis.

Results

We excluded non-native English speakers (n = 1), participants who reported being interrupted (n = 3) or in the presence of others while working on the experiment (n = 8), and participants who reported during the post-experimental questioning that in the memory test they were shown true alternatives to the false attributes they had seen previously (n = 11). We also excluded participants who exhibited incomplete data sets (n = 2). The following analysis is based on the remaining 225 participants. Including all participants in the analysis did not change the general pattern of results.

Memory of unpresented valid concepts. All memory probes associated with the experimental items were "new." The memory probes were expected to show a greater tendency to be confused as being seen previously (i.e., old) when the valid concept was accessed and processed during encoding. To assess whether the sentence type influenced the memory judgments of the unpresented valid concepts, we first conducted a one-way repeated-measures ANOVA on the factor Sentence Type (affirmative false vs. negated true vs. baseline). The responses on the 8-point memory scale served as the dependent variable. The analysis failed to support the hypothesis that the type of sentence matters in determining the judgments of memory of the valid concepts, $F(2, 448) = 1.49$, $p = .23$, $\eta_p^2 = .01$. The mean memory judgments of the valid concept following false sentences, negated sentences, and baseline sentences were close to each other (see Table 1).

Table 1: Means of memory for probe words as a function of Sentence Type. Ratings were made on an 8-point scale (1 = *new*, 8 = *old*). Numbers in parentheses depict the standard error for the statistical analysis.

Sentence type	*Memory probe*	
	valid concept (Experiment 1a)	false concept (Experiment 1b)
affirmative false	4.23 (.07)	5.44 (.07)
negated true	4.29 (.07)	5.31 (.07)
baseline	4.21 (.07)	

In addition, we compared the mean memory rating of the three types of target sentences (M = 4.24, SD = .94) to two benchmarks. First, these ratings were higher than the memory rating of "new" unrelated probes (M = 3.97, SD = 1.02), $F(1, 224) = 48.18$, $p < .001$, $\eta_p^2 = .18$, suggesting that the presence of the focal object in the sentence (e.g., "honey") increased the feeling that the related probe word (e.g., "bees") appeared in the original sentences. Second, these ratings were lower than ratings of "old" memory probes—those that refer to concepts actually presented in the sentences in the filler trials. These memory probes were judged as more "old" (M = 5.07, SD = .96) than were the valid concepts that had never been shown before, $F(1, 224) = 189.82$, $p < .001$, $\eta_p^2 = .46$.

To corroborate our findings, we also conducted linear mixed-model analyses using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in the statistical software R Version 3.4.1 for Windows (R Core Team, 2017). In the following we report the maximal random-effect structures that converged (Barr, Levy, Scheepers, & Tily, 2013). We fitted a model with Sentence Type (i.e., affirmative false vs. negated true vs. baseline) as fixed effect. The model had by-subject and by-item (i.e., nominal sentences used in the experiment; see Appendix) random intercepts, as well as by-item random slopes for Sentence Type. The analysis revealed a non-significant main effect of Sentence Type $\chi^2(2) = 3.78$, $p = .15$. Model parameters and estimates for the model are given in Table 2.

Table 2: Model parameters and estimates for the model, Experiment 1a.

Parameter	*Estimate*	*Std. Error*	*t-value*	*95% Confidence Interval*
intercept	4.24	0.10	43.51	4.05, 4.43
affirmative false vs. baseline	0.02	0.05	0.37	-0.08, 0.12
negated true vs. baseline	0.09	0.06	1.72	-0.01, 0.20

Because we could not reject the possibility that all three types of sentences had the same memory outcomes, our results taken together do not provide conclusive evidence that the presence of negation, or falsehood, influenced the encoding strength of the valid concept.

Performance in the grammar task. Might it be that the invariance in memory with respect to the sentence type reflects insensitivity to sentence type during encoding? In order to investigate whether participants were sensitive to the different sentence types, we performed a one-way repeated-measures ANOVA (Sentence Type: affirmative false vs. negated true vs. baseline), with the proportion of correct categorizations of sentences as "ordered" as the dependent variable. The analysis revealed a significant main effect of Sentence Type, $F(2, 448) = 15.49$, $p < .001$, $\eta_p^2 = .07$. To clarify this main effect, we computed contrasts comparing the different types of sentences to each other (see Table 3). These contrasts revealed that the rates of correct identification as "ordered" of true negated sentences and false affirmative sentences did not differ from each other, $F(1, 224) = 1.34$, $p = .25$, $\eta_p^2 = .01$. However, both types of sentences led to statistically inferior performance compared to the true baseline sentences: $F(1, 224) = 24.44$, $p < .001$, $\eta_p^2 = .10$ in the case of affirmative false; $F(1, 224) = 22.88$, $p < .001$, $\eta_p^2 = .09$ in the case of true negated. Thus, the falsehood of the sentences as well as the inclusion of negation seem to have somewhat interfered with the categorization task.

All filler trials involved sentences in a jumbled word order. Analyses of the filler trials revealed that categorizing sentences as "jumbled" was also affected by Sentence Type, $F(2, 448) = 53.20$, $p < .001$, $\eta_p^2 = .19$. Participants were most accurate for affirmative false sentences. Accuracy for these sentences was higher than for negated false sentences, $F(1, 224) = 52.86$, $p < .001$, $\eta_p^2 = .19$, and for affirmative true sentences , $F(1, 224) = 101.07$, $p < .001$, $\eta_p^2 = .31$. Thus, the falsehood of the sentence seemed to help the correct categorization of a sentence as "jumbled," with the exception being that negated false sentences were more difficult to categorize than affirmative false sentences, $F(1, 224) = 52.86$, $p < .001$, $\eta_p^2 = .19$. We are cautious in interpreting the accuracy outcomes of the fillers because unlike the experimental items, the type of filler sentence was confounded with the specific sentences that all participants saw.

Table 3: Means of proportion of correct categorization of sentences as ordered/jumbled as a function of Sentence Type in Experiments 1a and 1b. Numbers in parentheses depict the standard error for the statistical analysis.

Sentence Type	*Ordered*		*Sentence Type*	*Jumbled*	
	Experiment 1a	Experiment 1b		Experiment 1a	Experiment 1b
affirmative false	.94 (.01)	.91 (.01)	affirmative true	.86 (.01)	.84 (.01)
negated true	.95 (.01)	.93 (.01)	affirmative false	.92 (.01)	.90 (.01)
baseline	.97 (.01)	.95 (.01)	negated false	.88 (.01)	.86 (.01)

Discussion

The classification outcomes of the grammar task indicate a sentence-type effect: relative to the baseline, both negated sentences and false sentences interfered with detection of the sentences' grammatical order. However, the memory results from Experiment 1a suggest that encoding within a grammar task causes little difference in the encoding strength of the valid concept among the three types of sentences: affirmative false, negated true, and baseline. Yet, compared to completely new concepts, the valid concepts were more likely to

be confused as old, much in the same way as the related concepts presented in the baseline sentences.

One may attribute our findings to the shallow levels of semantic processing that the grammar task induces as compared to, for instance, asking participants questions about the sentence content (see Isberner & Richter, 2014). Accordingly, deeper semantic processing, beyond detecting a correct grammatical order, might be necessary to influence the encoding strength of the valid concept while processing false sentences and their negations, relative to baseline sentences. Deeper semantic processing might be induced by a task that has direct relevance for the processing of sentences' truth-values. We explore this possibility in Experiments 2a, 2b, and 3.

Although the results of Experiment 1a do not provide evidence for differences in the strength of encoding of the valid concept among false, negated, and baseline sentences, it is still possible that false and negated sentences differ with respect to the encoding strength of the false concept. Thus, Experiment 1b complements Experiment 1a by investigating the memory of the last word in the false and negated sentences, that is, the memory of the false concept.

Experiment 1b

Participants and Design

In line with Experiment 1a, we determined the sample size beforehand, with a total requirement of 250 participants. Slightly larger samples resulted from participants who took part in the experiment but did not request their compensation immediately after completing the study. If these participants asked for their compensation later, we granted it retroactively. We recruited 254 participants (144 female, 108 male, 1 other, 1 not reported; M_{age} = 39.02 years) via MTurk. We constrained participation to workers who did not take part in the first

experiment. All other recruiting criteria, payment, and the demographic post-experimental questionnaire were identical to Experiment 1a.

Procedure

The experimental design was identical to Experiment 1a with the exception of the memory task. The experimental probe words in the memory task consisted of the presented false concepts (e.g., "butterflies") that were shown in the affirmative false (e.g., "Honey is made by butterflies") and negated true sentences (e.g., "Honey is not made by butterflies") during the grammar task. The same probe word (e.g., "butterflies") was also shown following the baseline sentences in the grammar task (e.g., "Honey is made out of nectar"). However, because the probe words did not appear in the baseline sentences and because they were not directly related to the baseline sentences semantically, they are considered new. The probes related to the filler trials were changed to reflect the changes in the experimental trials: 20 probes related to affirmative false filler sentences (e.g., "Kids 'trick or treat' play on Christmas") and were valid concepts (e.g., "Halloween"). The other 40 probes were new words with no direct relation to the sentences shown during the grammar task.

Results

We excluded all non-native English speakers (n = 5), participants who reported being interrupted (n = 8) or in the presence of others while working on the experiment (n = 10), and those who indicated they had help solving the task (n = 2). In addition, we excluded one participant who indicated having dyslexia, and one participant who asked to be removed from the data set. One participant had an incomplete data set and was removed. The following analyses are based on the remaining 234 participants. Including all participants in the analyses did not change the general pattern of results.

Memory of presented false concepts. Both false and negated sentences were tested with the false concept, which was part of the sentence. Accordingly, an increasing tendency to rate the probe as "new" reflects a weaker encoding of the false concept. To assess whether

the sentence type had an influence on the recognition of false concepts, we conducted a repeated-measures ANOVA on the factor Sentence Type (affirmative false vs. negated true). The responses on the 8-point memory scale served as the dependent variable. We omitted probes related to the baseline sentences from this analysis because the memory probe word was never presented in the baseline sentence (see Appendix).

The analysis revealed a significant main effect for Sentence Type, $F(1, 233) = 6.25$, $p = .01$, $\eta_p^2 = .03$, indicating that participants judged the last word that was presented in the sentence as less old when the word had appeared in a negated sentence compared to when the word appeared in a false sentence (see Table 1). Categorization of concepts that were not shown in baseline sentences and should be considered as new ($M = 3.89$, $SD = 1.16$) differed significantly from affirmative false, $F(1, 233) = 374.14$, $p < .001$, $\eta_p^2 = .62$, and negated true sentences, $F(1, 233) = 396.29$, $p < .001$, $\eta_p^2 = .63$.

As in Experiment 1a, we analyzed the data with a mixed-model design using R. Specifically, we fitted a model with Sentence Type (i.e., affirmative false vs. negated true) as fixed effect, by-subject and by-item random intercepts, as well as by-item and by-subject random slopes for Sentence Type. The analysis revealed a significant main effect of Sentence Type, $\chi^2(1) = 7.67$, $p = .006$. Model parameters and estimates are given in Table 4.

Table 4: Model parameters and estimates for the model, Experiment 1b.

Parameter	*Estimate*	*Std. Error*	*t-value*	*95% Confidence Interval*
intercept	5.42	0.10	54.70	5.23, 5.62
affirmative false vs. negated true	-0.13	0.05	-2.84	-0.23, -0.04

Performance in the grammar task. We performed a one-way repeated-measures ANOVA on the results (Sentence Type: affirmative false vs. negated true vs. baseline), with

the proportion of correct classification of sentences as "ordered" as the dependent variable. The analysis revealed a significant main effect of Sentence Type, $F(2, 466) = 18.93$, $p < .001$, $\eta_p^2 = .08$. To clarify this main effect, we conducted contrast analyses comparing each level of Sentence Type to the other levels (see Table 3). These analyses revealed that the rate of correct identification as "ordered" of negated sentences and false sentences differed from each other, $F(1, 233) = 11.02$, $p = .001$, $\eta_p^2 = .05$. That is, participants were more accurate in classifying negated true sentences than affirmative false sentences. As in Experiment 1a, both these sentences led to inferior performance compared to the baseline sentences: $F(1, 233) = 27.10$, $p < .001$, $\eta_p^2 = .10$ in the case of affirmative false sentences; $F(1, 233) = 13.87$, $p < .001$, $\eta_p^2 = .06$ in the case of negated true sentences.

Analyses of the filler trials revealed that categorizing sentences as "jumbled" was also affected by the Sentence Type, $F(2, 466) = 39.71$, $p < .001$, $\eta_p^2 = .15$. Participants were most accurate for affirmative false sentences. Accuracy for these sentences was higher than for negated false sentences, $F(1, 233) = 24.94$, $p < .001$, $\eta_p^2 = .10$, and for affirmative true sentences, $F(1, 233) = 91.69$, $p < .001$, $\eta_p^2 = .28$. Thus, the falsehood of the sentence seemed to help the correct categorization of a sentence as "jumbled," with the exception being that negated false sentences were more difficult to categorize than affirmative false sentences, $F(1, 233) = 24.94$, $p < .001$, $\eta_p^2 = .10$.

Combined analysis, Experiments 1a and 1b. Although the assignment of participants to the two experiments was not random, we performed a combined analysis of Experiments 1a and 1b to examine whether the pattern of memory results for the valid and false concepts differs as a function of the type of sentence. A 2 (Memory Probe: valid unpresented concept vs. presented false concept) × 2 (Sentence Type: affirmative false vs. negated true) mixed-model ANOVA revealed a significant main effect of Memory Probe, $F(1, 457) = 136.99$, $p < .001$, $\eta_p^2 = .23$, which was qualified by a significant two-way interaction of Sentence Type and Memory Probe, $F(1, 457) = 7.33$, $p = .007$, $\eta_p^2 = .02$ (see Table 1). The pattern of means

indicates that memory of false concepts is weaker after processing a negated sentence than after processing a false sentence, while there was a non-significant difference in the other direction for the memory of valid concepts.

Discussion

Summing up the results from Experiments 1a and 1b, we found that under conditions in which the processing task had little relevance for semantic encoding of information and might have induced rather shallow levels of processing, there was an interaction between the type of memory probe and the type of sentence. That is, although we found no evidence that the encoding of the valid concepts differed between the three types of sentences, the memories of false concepts presented in the sentences were influenced by the sentence type, even in a relatively impoverished encoding task. Specifically, the false concepts showed a weaker presence in memory after respondents processed negated sentences as compared to the processing of false sentences. We believe that this might be explained by the impact of the explicit negation, namely, through negation-induced inhibition (e.g., MacDonald & Just, 1989). Based on Experiment 1b, we can infer that negation-induced inhibition might be triggered even when the task requires relatively shallow levels of semantic processing, and it occurs more strongly with explicit negations than after processing false sentences.

There are several alternative explanations that could account for our findings. Although the classification outcomes of the grammar task suggested that both negated sentences and false sentences interfered with detection of the sentences' grammatical order, we do not have direct evidence that respondents detected falsehood or processed the meaning of the negation. Experiments 2a and 2b address this.

Moreover, the Sentence-Type-by-Memory-Probe interaction we observed could be an artifact of sentence length, because sentences that contained a negation were longer by one word compared to false sentences. To wit, for the memory of presented words, the longer the sentence the worse should be memory of what was presented. Accordingly, memory of the

false concepts was worse following negated sentences than following false sentences. For the memory of valid concepts, such sentence-length effects should not hold. Our data are consistent with the suggestion of no difference among the three sentence types in the memory of the valid concepts. This might indicate that participants indeed did not process the full sentences. Experiments 2a and 2b are less amenable to these possibilities because they employ a task that requires more meaningful encoding of the stimulus information. We hypothesize that when the task induces more elaborative encoding, differences among negated true, affirmative false, and baseline sentences should be amplified, allowing us to rule out alternative explanations.

Experiment 2a

To induce more meaningful processing of information, we employed an impression-formation task. Impression formation requires integration (Burnstein & Schul, 1982, 1983) and induces deep levels of encoding (Craik, 2002; Craik & Tulving, 1975). Briefly, when people form impressions they go beyond the information given in the communication and activate inferences to support the goals that the communication and context afford (Asch, 1946; Heider, 1944; see relevant discussions in Wyer & Srull, 1986). To this end, in Experiments 2a and 2b, true and false sentences were paired with pictures of persons, and participants' task was to form impressions of these persons while taking these sentences into account.

In the context of impression formation, reception of falsehoods highlights to the recipients the truthfulness of the source and the danger of being misled. People expect sources of information to be truthful (Grice, 1975), and differentiating between sources that speak the truth and those that provide falsehoods helps confirm or disconfirm this expectation (Schul et al., 2004). A great deal of research shows that the trustworthiness of the source is of critical

importance when people evaluate messages (e.g., Brinol & Petty, 2009; Chaiken & Maheswaran, 1994; Hovland & Weiss, 1951; Smith et al., 2013; Sparks & Rapp, 2011), and that recipients' attention to source-related information is influenced by the plausibility of the information the source provides (Braasch, Rouet, Vibert, & Britt, 2012). Therefore, when a source makes a blatantly false assertion (e.g., "honey is made by butterflies"), the concern with truth dictates that the recipient should pay attention to the false communication and consider it an indication about the source itself. This should result in two outcomes. First, the evaluation of sources in an impression-formation task should reflect whether falsehood or truth was detected. Specifically, sources associated with false sentences should be evaluated negatively, and those associated with true sentences should be viewed positively (see Hughes, Ye, Van Dessel, & De Houwer, 2018; Unkelbach, Bayer, Alves, Koch, & Stahl, 2011). Thus, the outcomes of the impression-formation task allow us to see whether falsehood has been detected.

Second, sources' trustworthiness should influence the encoding strength of false and valid concepts. When a source of information negates falsehood, he or she might be viewed as trustworthy. When recipients encounter trustworthy sources, they are less likely to be concerned with the exact phrasing of the communication or with the characteristics of the source. Rather, they become more concerned with clarity of the message (Sachs, 1974), tending to focus more on the gist of the information and less on the exact details. As a result, recipients of negated sentences may shift processing from the complex phrasing that includes the negation of the false concept to the simplified phrasing that includes the valid concept (Mayo et al., 2004). Accordingly, there should be a tendency to encode the negation in terms of the valid concept (see Fillenbaum, 1966) rather than the false concept.

When the source of communication provides a false sentence, he or she becomes untrustworthy, and the relative encoding of false and valid concepts might be influenced in two opposite ways. First, to protect themselves from the danger of receiving false

information, recipients may prioritize the encoding of valid concepts (see Schul et al., 2004). Alternatively, the danger of being misled might result in an encoding priority of false concepts because maintaining falsehood, rather than correcting falsehood, might help one remember that a source was untruthful, thus influencing their impression, and allowing recipients to protect themselves from the source in future communications (Hovland & Weiss, 1951).

Thus, Experiments 2a and 2b explored whether negated and false sentences differ with respect to the relative encoding strength of false and valid concepts when the truth-value of sentences has important implications for the task at hand.

Participants and Design

In line with our previous experiments, we determined the sample size beforehand, with a total requirement of 250 participants. Slightly larger samples resulted from participants who took part in the experiment but did not request their compensation immediately after completing the study. If these participants asked for their compensation later, we granted it retroactively. We recruited 256 participants (160 female, 96 male; M_{age} = 37.93 years) via MTurk. We constrained participation to workers who did not take part in the first two experiments. All other recruiting criteria, payment, and the demographic post-experimental questionnaire[2] were identical to Experiments 1a and 1b. The experimental design was identical to Experiment 1a with the exception of the impression-formation task.

Procedure

The procedure and the stimulus material were identical to Experiment 1a, except for one critical difference. Sentences were presented as part of an impression-formation task in which we requested participants to form impressions of male persons. We instructed them as

[2] The post-experimental questionnaire in Experiment 2a had an additional question asking whether participants followed the instructions in the first part of the experiment by paying attention to both the presented sentence and the person presented together with the sentence.

follows: “Research has shown that we form impressions about people easily, only based on their appearance. In our research we want to investigate how a statement made by a person can influence impression formation. Therefore, it is very important that you base your impression on both the picture and the statement.” In each trial, participants were first shown a sentence (that was true or false in an affirmative or negated phrasing). The sentence appeared centered in the upper part of the screen and was shown alone for 2000 ms. It was followed by a black-and-white photo of a male person, in side view, presented under the sentence in a centered position (see Figure 2). Black-and-white photos were taken from the Face Recognition Technology database (FERET; Phillips, Moon, Rizvi, & Rauss, 2000; Phillips, Wechsler, Huang, & Rauss, 1998). We asked participants to report their feeling toward the person, based on his picture and based on what he said. Participants pressed the “p” key on their keyboard to indicate they had an overall good feeling and pressed the “q” key to indicate an overall bad feeling. The sentence and the photo both stayed on the screen until participants indicated their answer.

Each participant saw 120 different sentences, 60 of which were experimental trials. The filler sentences were identical to the filler sentences in Experiment 1a with the exception that they were shown in correct grammatical order. Each of the 120 sentences was paired with a different photo. All participants saw the same pairing of sentences and photos across conditions. In particular, the same photo appeared irrespective of whether an experimental sentence was affirmative false, negated true, or baseline sentence.

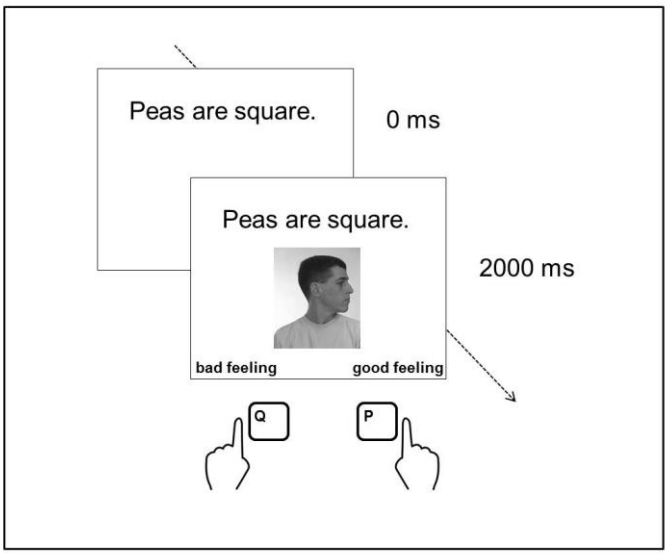

Figure 2: Example of trial sequence of the impression-formation task in Experiments 2a and 2b.

Results

We excluded all non-native English speakers (n = 2), those who reported being interrupted (n = 5) or being in the presence of others while working on the experiment (n = 8), and those who realized that they were shown in the memory test true alternatives to the false attributes they had previously seen (n = 6). The following analysis is based on the remaining 236 participants. Including all participants in the analysis did not change the general pattern of results.

Memory of the unpresented valid concepts. We analyzed the memory scores in a one-way repeated-measures ANOVA. The analysis revealed a significant main effect for Sentence Type, $F(2, 470) = 3.70$, $p = .026$, $\eta_p^2 = .02$ (see Table 5). Contrast analyses indicated that probes were rated as similarly old after participants processed false sentences and baseline sentences, $F(1, 235) = .03$, $p = .86$, $\eta_p^2 = .00$. Importantly, the memory probes that referred to the negated sentences were rated as more "old." Specifically, after processing negations, participants rated probes as more old in comparison to the baseline, $F(1, 235) = 6.07$, $p = .01$, $\eta_p^2 = .03$, and to the false sentences, $F(1, 235) = 5.06$, $p = .025$, $\eta_p^2 = .02$.

Probe words that were taken from the filler trials (i.e., "old" probes) were rated as significantly more old (M = 5.86, SD = .83) than all other probe words (*all* $p < .001$). Memory ratings of probe words that were completely new, that is, probe words with no direct relation to any of the presented sentences, were significantly below the memory scores for the three types of experimental sentences (M = 3.85, SD = 1.11; *all* $p < .001$).

Table 5: Means of memory for probe words as a function of Sentence Type. Ratings were provided on an 8-point scale (1 = *new*, 8 = *old*). Numbers in parentheses depict the standard error for the statistical analysis.

Sentence type	*Memory probe*	
	valid concept (Experiment 2a)	false concept (Experiment 2b)
affirmative false	4.09 (.07)	5.45 (.08)
negated true	4.21 (.07)	5.18 (.07)
baseline	4.08 (.07)	

We also analyzed the data with a mixed-model design using R. Specifically, we fitted a model with Sentence Type (i.e., affirmative false vs. negated true vs. baseline) as fixed effect and by-subject and by-item random intercepts, as well as by-item and by-subject random slopes for Sentence Type. The analysis revealed a significant main effect of Sentence Type $\chi^2(2) = 7.67$, $p = .02$. Model parameters and estimates for the model are given in Table 6.

Table 6: Model parameters and estimates for the model, Experiment 2a.

Parameter	*Estimate*	*Std. Error*	*t-value*	*95% Confidence Interval*
intercept	4.07	0.10	42.75	3.89, 4.26
affirmative false vs. baseline	0.01	0.05	0.12	-0.09, 0.10
negated true vs. baseline	0.13	0.05	2.51	0.03, 0.23

Impression formation. The impressions in the experimental trials varied as a function of the variant of the sentence. For each person and each type of sentence (affirmative false vs. negated true vs. baseline), we computed the proportion of *good feeling* responses toward the person. We conducted a one-way ANOVA, with Sentence Type as a repeated-measures factor, on these proportions. The analysis revealed a significant main effect for Sentence Type, $F(2, 470) = 536.39$, $p < .001$, $\eta_p^2 = .70$ (see Figure 3). To clarify this main effect, we tested the pairwise contrasts comparing the levels of Sentence Type. These analyses revealed that false sentences ($M = .24$, $SD = .23$) led to fewer positive evaluations than negated true sentences ($M = .71$, $SD = .21$), $F(1, 235) = 491.60$, $p < .001$, $\eta_p^2 = .68$; and negated sentences led to fewer positive evaluations than baseline sentences ($M = .80$, $SD = .20$), $F(1, 235) = 73.38$, $p < .001$, $\eta_p^2 = .24$. Thus, participants strongly disliked sources that provided false sentences and liked sources that used negated phrases somewhat less than sources that used affirmative phrases. This finding indicates that in forming their impressions, participants were sensitive to the type of sentence.

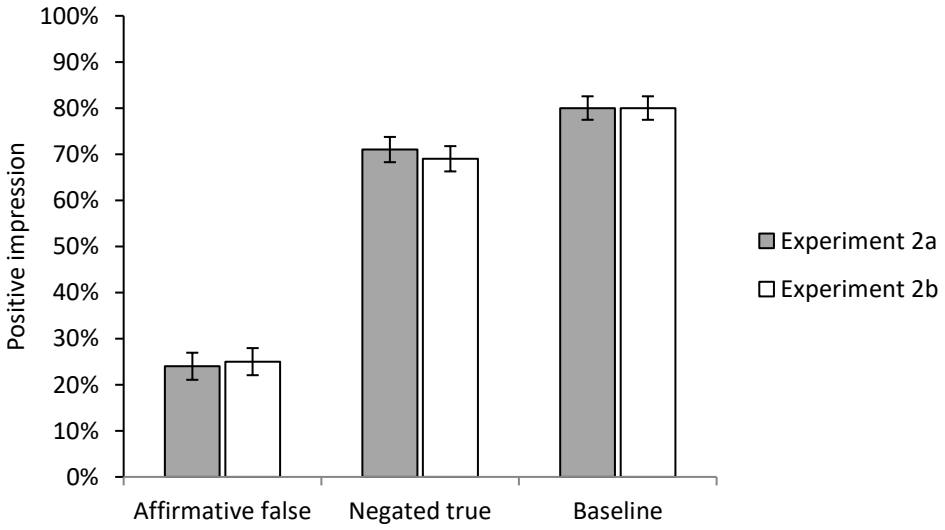

Figure 3: Mean percentages of positive impression as a function of Sentence Type for Experiments 2a and 2b. Error bars depict 95% confidence intervals.

Discussion

The impression ratings differed for the three types of sentences. That is, false sentences triggered mostly negative impressions, and negated sentences elicited mostly positive impressions; however, the latter impressions were less positive than those resulting from the baseline sentences. The memory outcomes followed a different pattern: Negated sentences led to stronger encoding of the valid concept than did false sentences. In fact, we found no evidence that the valid concept was encoded any more strongly following the false sentences compared to baseline sentences. Experiment 2b complements Experiment 2a by investigating the encoding of the false concept that was shown within the sentence during the impression-formation task.

Experiment 2b

Participants and Design

In line with previous experiments, we determined the sample size beforehand, with a total requirement of 250 participants. Slightly larger samples resulted from participants who took part in the experiment but did not request their compensation immediately after completing the study. If these participants asked for their compensation later, we granted it retroactively. We recruited 267 participants (160 female, 100 male, 7 not reported; M_{age} = 36.63 years) via MTurk. We constrained participation to workers who did not take part in the first three experiments. All recruiting criteria, payment, and the demographic post-experimental questionnaire[3] were identical to the previous experiments. The experimental

[3] The post-experimental questionnaire in Experiment 2b had an additional question asking whether participants followed the instructions in the first part of the experiment by paying attention to both the presented sentence and the person presented together with the sentence. The post-questionnaire also contained a six-item dispositional trust questionnaire (Yamagishi & Yamagishi, 1994).

design was identical to Experiment 2a with the exception of the memory test. The memory test in Experiment 2b was identical to the memory test in Experiment 1b.

Results

We excluded all non-native English speakers (n = 2), those who reported being interrupted (n = 3) or being in the presence of others while working on the experiment (n = 9), and one participant who indicated in the post-experimental questionnaire, apparently due to the filler trials, that he/she was shown true alternatives to the false attributes viewed previously. In addition, we excluded participants who exhibited incomplete data sets (n = 7). The following analyses are based on the remaining 245 participants. Including all participants in the analyses did not change the general pattern of results.

Memory of presented false concepts. We conducted a one-way repeated-measures ANOVA on the factor Sentence Type (affirmative false vs. negated true). The analysis revealed a significant main effect for Sentence Type, $F(1, 244) = 19.94$, $p < .001$, $\eta_p^2 = .08$ (see Table 4). Replicating the results of Experiment 1b, participants rated a word that actually appeared in the sentences of the impression-formation phase as less old when it appeared in a negated sentence than when it appeared in a false sentence. Recognition of concepts that were not shown in baseline sentences and should be considered as new (M = 3.67, SD = 1.31) differed significantly from affirmative false, $F(1, 244) = 389.53$, $p < .001$, $\eta_p^2 = .62$, and negated true sentences, $F(1, 244) = 338.94$, $p < .001$, $\eta_p^2 = .58$.

We also performed a mixed-model analysis, with Sentence Type (i.e., affirmative false vs. negated true) as fixed effect and by-subject and by-item random intercepts, as well as by-item and by-subject random slopes for Sentence Type. The analysis revealed a significant main effect of Sentence Type, $\chi^2(1) = 18.63$, $p < .001$. Model parameters and estimates for the model are given in Table 7.

Table 7: Model parameters and estimates for the model, Experiment 2b.

Parameter	*Estimate*	*Std. Error*	*t-value*	*95% Confidence Interval*
intercept	5.42	0.12	46.18	5.19, 5.65
affirmative false vs. negated true	-0.25	0.06	-4.60	-0.36, -0.14

Impression formation. The impression-formation results were virtually identical to those in Experiment 2a (Figure 3). To assess whether the sentence type had an influence on impression formation, we conducted a one-way repeated-measures ANOVA on the factor Sentence Type (affirmative false vs. negated true vs. baseline). We found a significant main effect for Sentence Type, $F(2, 488) = 472.03$, $p < .001$, $\eta_p^2 = .66$. To clarify this main effect, we computed contrasts comparing each level of Sentence Type. These revealed that affirmative false sentences ($M = .25$, $SD = .24$) led to fewer positive impressions than negated true sentences did ($M = .69$, $SD = .22$), $F(1, 244) = 418.88$, $p < .001$, $\eta_p^2 = .63$, and negated true sentences led to fewer positive impressions than (true) baseline sentences did ($M = .80$, $SD = .20$), $F(1, 244) = 91.01$, $p < .001$, $\eta_p^2 = .27$, thus reinforcing the claim that in forming their impressions, participants were sensitive to the type of sentence.

Combined analysis, Experiments 2a and 2b. In spite of the non-random assignment of participants to experiments, we compared the memory outcome in the two experiments using a 2 (Sentence Type: affirmative false vs. negated true) × 2 (Memory Probe: valid alternative vs. false presented) mixed-model ANOVA. The analysis revealed a significant main effect of Memory Probe, $F(1, 479) = 144.70$, $p < .001$, $\eta_p^2 = .23$, which was qualified by a significant two-way interaction of Sentence Type and Memory Probe, $F(1, 479) = 23.58$, $p < .001$, $\eta_p^2 = .05$ (see Table 5). The interaction pattern indicates that memory for false concepts was weaker after processing negated true sentences than after processing affirmative false sentences, while memory for valid concepts was stronger after negated true sentences than after affirmative false sentences.

Discussion

Summing up the results from Experiments 2a and 2b, we found that the positivity of the impression formation was influenced by the type of sentence. Sources who were associated with false sentences were evaluated less favorably than sources who were associated with true sentences. Thus, we can assume that falsehoods were detected and negations were processed. The memory data showed that false and negated sentences differed with respect to the strength of encoding of valid and false concepts when they were encoded within an impression-formation task. In particular, negated sentences were associated with stronger encoding of the valid concept and weaker encoding of the false concept relative to false sentences. The results rule out the possibility that sentence length is an alternative explanation for our findings, because valid concepts led to stronger encoding for negated sentences as compared to false sentences and baseline sentences. This can only occur when the full sentences, containing negations, are processed. The pattern of results is consistent with the hypothesis that forming an impression about untruthful sources leads to an encoding priority of false concepts, while forming an impression of truthful sources leads to an encoding priority of valid concepts.

Experiment 3 attempts a more fine-grained analysis of the impression-formation task. Our operationalization of impression formation involved two components—the presence of a source and an explicit evaluation task. Participants were asked to report their feeling toward sources, taking sentences that sources made into account. Experiment 3 investigates whether we can attribute the effects found in Experiments 2a and 2b to the presence of the source or to the evaluation task itself. To this end, half of the participants in Experiment 3 evaluated sources and their statements on a good–bad dimension (as in Experiments 2a and 2b), and the other half evaluated only the statements on a good–bad dimension.

Experiment 3 addresses a methodological limitation of the previous experiments as well. In Experiments 1a, 1b, 2a, and 2b, participants rated the experimental sentences with

respect to either the memory of the valid unpresented concept (1a, 2a) or the memory of the presented false concept (1b, 2b), making the memory probe a factor that varied between experiments. However, as suggested by the combined analyses, our main argument is informed by the interaction between valid and false concepts as a function of Sentence Type. Thus, Experiment 3 manipulated the factor Memory Probe within-participant.

Experiment 3

Participants and Design

In line with previous experiments and to account for an additional between-subjects factor, we determined the sample size beforehand, with a total requirement of 500 participants. Slightly larger samples resulted from participants who took part in the experiment but did not request their compensation immediately after completing the study. If these participants asked for their compensation later, we granted it retroactively. We recruited 519 participants (271 female, 229 male, 6 other, 13 unknown; M_{age} = 34.40 years) via Prolific Academic (e.g., Peer, Brandimarte, Samat, & Acquisti, 2017). Participants could sign up for the experiment only if they were United States nationals, currently residing in the United States, and if they were native English speakers. We required that they held a record of supplying acceptable data at least 90% of the time. They received £1.67 (approx. $1.30) for their participation. The post-experimental demographic questionnaire was identical to the previous experiments. Data collection was interrupted once due to technical issues.

The study consisted of a 3 (Sentence Type: affirmative false vs. negated true vs. baseline true) × 2 (Memory Probe: old vs. new alternative) × 2 (Task: person impression vs. sentence impression) mixed-model design, with the first two factors being manipulated within-participant and the third being a between-participants factor.

Procedure

Similar to Experiments 2a and 2b, Experiment 3 consisted of three consecutive phases: an impression-formation phase, a filler task, and a memory-test phase. The *person-impression* condition was identical to the impression-formation phase in Experiments 2a and 2b. In the *sentence-impression* condition, participants were presented only with a sentence and were instructed to indicate whether they have an overall good or bad feeling toward the presented sentence. The instructions were as follows: "First you will be presented with a statement. Please read this statement carefully. Then you will be asked to indicate your overall feeling toward the statement. Please indicate whether you have an overall good or bad feeling toward the statement." The sentence appeared centered in the upper part of the screen and was shown for 2000 ms. We asked participants to report their feeling toward the sentence by pressing the "p" key on their keyboard to indicate they had an overall good feeling, and pressing the "q" key to indicate an overall bad feeling. Participants' responses were recorded only after 2000 ms. The sentence stayed on the screen until participants indicated their answer. The subsequent filler task was identical to that of the previous experiments.

The type of memory probe in the memory-test phase was manipulated within-participant. Specifically, each participant was presented with 90 experimental memory-probe words (plus 10 fillers). The probe words included 10 valid concepts (i.e., "new" memory probes, not presented in the sentences) for negated sentences and 10 valid concepts for false sentences, as well as 10 false concepts (i.e., "old" memory probes, presented within the sentences) for negated sentences and 10 false concepts for false sentences. Memory of probes from 20 baseline sentences was also tested with the valid or the false concept. However, because valid and false concepts were not included in the baseline sentences, all 20 memory probes relevant to the baseline sentences are considered new. Finally, we included 30 memory probes that were completely new, namely, unrelated to any of the sentences. We counterbalanced (between-participants) which nominal sentences served as reference for the old or new memory probes. This resulted in four different counterbalancing conditions.

Each participant started the memory task with the same 10 filler probes (6 old words, 4 new words), which were not included in the analysis. In contrast to the previously reported experiments, there were no additional fillers. Accordingly, each participant saw a total of 100 memory probes during the memory-test phase.

Results

We excluded all non-native English speakers (n = 8), participants who reported being interrupted (n = 7) or in the presence of others while working on the experiment (n = 17), those who indicated they had help solving the task (n = 1) or wrote down the key assignment instead of memorizing it (n = 13), and those who indicated in the post-experimental questionnaire that during the memory test they were shown true alternatives to the false attributes they had previously seen (n = 11). In addition, we excluded one participant who indicated that he or she confused the keys in most parts of the experiment, and one participant who indicated strongly disliking the experiment. Data sets of 17 participants were incomplete and were removed. Accordingly, the following analyses are based on the remaining 448 participants. Including all participants in the analyses did not change the general pattern of results.

Memory of false and valid concepts. We start by comparing the memory judgments for the valid and false concepts. The memory scores were analyzed in a 2 (Sentence Type: affirmative false vs. negated true) × 2 (Memory Probe: valid alternative vs. false presented) × 2 (Task: person impression vs. sentence impression) mixed-model ANOVA. Memory ratings of probes related to the baseline sentences were omitted from this analysis.

The analysis revealed a significant main effect for Memory Probe, $F(1, 446) = 644.18$, $p < .001$, $\eta_p^2 = .59$, which was qualified by a significant interaction of Memory Probe and Sentence Type, $F(1, 446) = 23.02$, $p < .001$, $\eta_p^2 = .05$ (see Table 8). To clarify this interaction, we compared the two types of sentences for valid-concept and false-concept memory probes. The valid-concept probes were judged as older after negated sentences (M = 4.06, SD = 1.36)

than after false sentences (M = 3.94, SD = 1.37), $F(1, 447) = 6.26$, $p = .01$, $\eta_p^2 = .01$, suggesting that the valid concepts were encoded more strongly after negated true sentences than after affirmative false sentences. The false concepts were rated as less old when they were embedded in negated sentences (M = 5.33, SD = 1.22) than when they appeared in false sentences (M = 5.55, SD = 1.14), $F(1, 447) = 14.77$, $p < .001$, $\eta_p^2 = .03$, suggesting that the false concepts were encoded less strongly after negated true sentences than after affirmative false sentences. Thus, Experiment 3 replicates the pattern of findings observed in Experiments 2a and 2b. The three-way interaction of Memory Probe, Sentence Type, and Task was not significant, $F(1, 446) = 2.74$, $p = .10$, $\eta_p^2 = .01$, a finding consistent with the observation that the 2 (Sentence Type) × 2 (Memory Probe) interaction pattern is not very different in the sentence-impression and the person-impression conditions. Thus, we did not find evidence for the importance of the source of information for the pattern of memory of the valid and false concepts. That is, the evaluation task itself rather than the presence of a source seems to be critical for the obtained effects.

Notwithstanding, the analysis revealed that the presence (vs. absence) of a source led to a significant main effect for Task, $F(1, 446) = 5.94$, $p = .02$, $\eta_p^2 = .01$, which was qualified by a significant interaction of Task and Memory Probe, $F(1, 446) = 18.46$, $p < .001$, $\eta_p^2 = .04$. The pattern of these findings indicates that, overall, valid concepts were encoded more strongly in the person-impression condition (M = 4.22, SD = 1.22) as compared to the sentence-impression condition (M = 3.75, SD = 1.26), while there was virtually no difference in the memory of the false concept (M = 5.43, SD = 1.05 vs. M = 5.46, SD = .97). We can only speculate that this pattern occurs because sources served as additional information that needed to be processed and integrated with the sentence information. As a result, they made it easier to distinguish between actually presented and new information in the sentence-impression condition.

Our next analysis focuses on the valid concepts. It tests whether negated and false sentences lead to an increase in encoding of valid concepts relative to baseline level. To this end, the memory judgments of the valid concepts were analyzed in a 3 (Sentence Type: affirmative false vs. negated true vs. baseline) × 2 (Task: person impression vs. sentence impression) mixed-model ANOVA. The analysis revealed a non-significant two-way interaction of Sentence Type and Task, $F(2, 892) = 1.43$, $p = .24$, $\eta_p^2 = .00$, suggesting that we do not have evidence supporting a Sentence Type difference between the person-impression and the sentence-impression tasks. Accordingly, it is meaningful to examine the main effect of Sentence Type, $F(2, 892) = 6.61$, $p = .001$, $\eta_p^2 = .02$ (see Table 8). Contrast analyses showed that, as in Experiment 2a, probes' memory reports were similar after participants processed false sentences ($M = 3.94$, $SD = 1.37$) and baseline sentences ($M = 3.89$, $SD = 1.34$), $F(1, 447) = .96$, $p = .33$, $\eta_p^2 = .00$. However, the memory probes that corresponded to the negated sentences ($M = 4.06$, $SD = 1.36$) were rated as older than the probes corresponding to the two other sentence types. Specifically, following negated sentences, participants rated probes as older compared to the baseline, $F(1, 447) = 11.46$, $p = .001$, $\eta_p^2 = .03$, and to the false sentences, $F(1, 447) = 6.26$, $p = .013$, $\eta_p^2 = .01$.

Finally, memory ratings of probe words that were completely new—that is, probe words with no direct relation to any of the presented sentences—were significantly below the memory scores for the three types of experimental sentences ($M = 3.20$, $SD = 1.21$; *all* $p < .001$).

Table 8: Means of memory for probe words as a function of Sentence Type and Task for Experiment 3. Ratings were made on an 8-point scale (1 = *new*, 8 = *old*). Numbers in parentheses depict the standard error for the statistical analysis.

Sentence type	*Memory probe*

	Person impression		Sentence impression	
	valid concept	false concept	valid concept	false concept
affirmative false	4.19 (.09)	5.52 (.08)	3.65 (.09)	5.59 (.08)
negated true	4.25 (.09)	5.34 (.08)	3.85 (.09)	5.32 (.09)
baseline	4.08 (.09)		3.67 (.09)	

We used a mixed-model analysis to fit a model with Sentence Type (i.e., affirmative false vs. negated true), Memory Probe (i.e., valid vs. false), and Task (i.e., person impression vs. sentence impression) as fixed effects. The model had by-subject and by-item random intercepts, as well as by-item and by-subject random slopes for Sentence Type. The analysis revealed a significant interaction of Memory Probe and Sentence Type, $\chi^2(1) = 26.67$, $p < .001$, and a non-significant interaction of Memory Probe, Sentence Type, and Task, $\chi^2(1) = 2.05$, $p = .15$. Model parameters and estimates for the model omitting the three-way interaction but including all two-way interactions are given in Table 9.

Table 9: Model parameters and estimates for the model, Experiment 3.

Parameter	*Estimate*	*Std. Error*	*t-value*	*95% Confidence Interval*
Intercept	4.62	0.16	28.17	4.30, 4.94
Sentence Type	0.10	0.12	0.81	-0.13, 0.33
Memory Probe	0.90	0.11	8.24	0.69, 1.12
Task	-0.47	0.10	-4.83	-0.66, -0.28
Sentence Type × Memory Probe	-0.35	0.07	-5.17	-0.48, -0.22
Sentence Type × Task	0.01	0.07	0.17	-0.13, 0.15
Memory Probe × Task	0.49	0.07	7.31	0.36, 0.62

The impression-formation tasks. A 3 (Sentence Type: affirmative false vs. negated true vs. baseline true) × 2 (Task: person impression vs. sentence impression) mixed-model ANOVA revealed a significant main effect of Task, $F(1, 446) = 18.51$, $p < .001$, $\eta_p^2 = .04$, and a significant main effect of Sentence Type, $F(2, 892) = 1876.07$, $p < .001$, $\eta_p^2 = .81$,

qualified by a significant interaction of Task and Sentence Type, $F(2, 892) = 67.53$, $p < .001$, $\eta_p^2 = .13$. To clarify this interaction, we computed the simple effects contrasts comparing each level of Sentence Type within the person-impression condition and the sentence-impression condition. As can be seen in Figure 4, the three types of sentence differ from each other within each of the two tasks, but the impressions were more extreme in the sentence-impression condition.

Specifically, in the person-impression condition, persons were evaluated more favorably when they were shown with a negated true sentence ($M = .69$, $SD = .21$) than when they were shown with an affirmative false sentence ($M = .22$, $SD = .22$), $F(1, 239) = 526.17$, $p < .001$, $\eta_p^2 = .69$. Moreover, persons were evaluated more favorably when they were shown with a true baseline sentence ($M = .77$, $SD = .21$) than when they were shown with a negated true sentence, $F(1, 239) = 58.99$, $p < .001$, $\eta_p^2 = .20$, or with an affirmative false sentence, $F(1, 239) = 598.55$, $p < .001$, $\eta_p^2 = .72$.

The sentence-impression condition showed a similar yet stronger pattern of effects, as revealed by the abovementioned interaction. Sentences were evaluated more favorably when they were negated true sentences ($M = .79$, $SD = .17$) than when they were affirmative false sentences ($M = .11$, $SD = .14$), $F(1, 207) = 1438.15$, $p < .001$, $\eta_p^2 = .87$. Moreover, sentences were evaluated more favorably when they were true baseline sentences ($M = .92$, $SD = .08$) than when they were negated true sentences, $F(1, 207) = 148.78$, $p < .001$, $\eta_p^2 = .42$, or when they were affirmative false sentences, $F(1, 207) = 4125.51$, $p < .001$, $\eta_p^2 = .95$. This suggests that participants were sensitive to the type of sentence as well as the presence of the source, and it appears that the information about the source (i.e., the picture of that person) diluted the sentence effect.

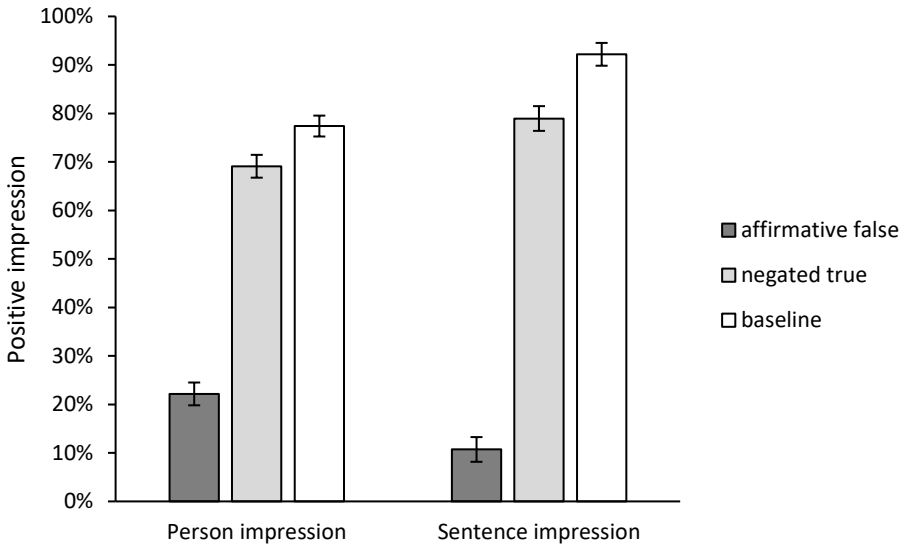

Figure 4: Mean percentages of positive impression as a function of Sentence Type for Experiment 3. Error bars depict 95% confidence intervals.

Combined Analysis of the Effects of the Three Encoding Tasks

Because the Sentence Type × Memory Probe interaction was significant in all three processing tasks—namely grammar, person impression, and sentence impression—we examined statistically whether our findings imply that the three tasks influence the Sentence Type × Memory Probe interaction differently. To do so, we combined the data from all five experiments (1a, 1b, 2a, 2b, and 3), with the question of interest being whether the task interacted with the pattern of findings reported in the different experiments. To anticipate the conclusion, we do not have statistical evidence to support the claim that the two-way interaction between Sentence Type and Memory Probe differs as a function of the processing task.

Specifically, our analyses focused on two interactions: the two-way interaction of Sentence Type (affirmative false vs. negated true) and Memory Probe (valid concept vs. false concept), and the three-way interaction of Sentence Type, Memory Probe, and Experiment (1a/b, 2a/b, 3). The two-way interaction compares the encoding of the false and valid

concepts for negated true and affirmative false sentences in all reported experiments. The three-way interaction examines whether the effect differs for the three separate experiments.

In order to examine the former, we fitted a model including the two-way interaction of Sentence Type and Memory Probe as fixed effect, controlling for all main effects. The model featured by-subject and by-item random intercepts, as well as a by-subject and by-item random slope for Memory Probe. The analysis revealed a significant two-way interaction of Sentence Type and Memory Probe, $\chi^2(1) = 61.02$, $p < .001$, indicating that averaging over the processing tasks, encoding strength of valid concepts was more pronounced for negated true sentences as compared to affirmative false sentences, and encoding strength of false concepts was less pronounced for negated true sentences as compared to affirmative false sentences.

Next, we tested whether this two-way interaction differs for the three tasks. In order to investigate this question, we fitted a model including the three-way interaction of Sentence Type (affirmative false vs. negated true), Memory Probe (valid concept vs. false concept), and Experiment (1a/b, 2a/b, 3) as fixed effect, while controlling for all two-way interactions and main effects. The model had by-subject and by-item random intercepts, as well as a by-subject and by-item random slope for Memory Probe. The analysis revealed a non-significant three-way interaction of Sentence Type, Memory Probe, and Experiment, $\chi^2(1) = 2.74$, $p = .09$.

Thus, taken together, the statistical analyses provide strong evidence that the memory of the valid and false concepts differs as a function of the type of sentence. Moreover, our findings fail to reveal strong evidence that the type of encoding matters for this interaction. Our findings are consistent with the hypothesis that across the three tasks, which are very different from each other, the valid concept is triggered more strongly by negated sentences than by false sentences; and at the same time, the false concept is more strongly triggered by false sentences than by negated sentences. This could be taken as a testimony for the enhanced encoding of the false information, which is consistent with the literature on reliance

on misinformation and belief perseverance (e.g., Rapp & Braasch, 2014; Schul & Burnstein, 1998), demonstrating that people tend to be influenced by false information even when they know this information is false.

General Discussion

The danger of receiving false information is omnipresent, and people might be highly vigilant against being influenced by falsehoods. Yet, as research on misinformation reveals, people are often biased by false information, even when they know the correct answer. The question is why? Our research provides an answer to this riddle: Even when people reject a sentence as false, they tend to focus on the false concept rather than on the valid concept. But, in order to mentally correct a false sentence, recipients need to consider the valid answer. When we compared two triggers for the correction of falsehood—a sentence consisting of clearly false information and a sentence consisting of an explicit negation of this information—we found that the valid concept exhibits a weaker presence in memory, and the false concept a stronger presence in memory, following the comprehension of false information as compared to its negation. In short, although evidently false, these sentences bias memory away from the truth.

Experiment 1a investigated the encoding of the valid concept, within a grammar task that was irrelevant to the sentences' truth-value and might have induced rather shallow processing levels. The statistical analysis failed to support encoding differences of valid concepts between negated and false information, leading us to investigate tasks that require more meaningful processing (e.g., Isberner & Richter, 2014; Nieuwland & Kuperberg, 2008). Notwithstanding, Experiment 1b demonstrates that even such shallow processing was sufficient to reduce the strength of encoding of the negated concept, in line with predictions

of negation-induced-inhibition models (e.g., de Vega et al., 2016; Kaup, 2001; MacDonald & Just, 1989). Together, this pattern suggests that a negation of falsehood is less likely to result in an encoding of false concepts than obvious false information, even when elaborative encoding is minimal.

Experiments 2a and 2b tested whether the pattern of encoding of the valid and the false concepts varies when processing is more meaningful and takes place within an impression-formation task in which the truth-value of sentences is of importance. We found that more meaningful processing induced more pronounced encoding effects. Specifically, negated falsehoods were associated with an enhanced encoding of the valid concept and a weaker encoding of the false concept as compared to obvious false sentences. Experiment 3 investigated whether the presence of a source is critical for this pattern of encoding effects. The findings suggest that the evaluation task is responsible for the abovementioned effects.

Why Are Falsehoods Not Corrected Unless They Are Explicitly Negated?

The difference between negated and false messages with respect to the encoding of false and valid concepts might be explained by the pragmatic implications attributed to negation. Explicit negation might be interpreted as a "refute" cue, previously addressed in persuasion research (e.g., Schul & Mazursky, 1990), leading recipients to inhibit the false concept and encode the valid concept. For reasons discussed below, false information might not trigger such a "refute" response, and so the difference between false information and negated information might have to do with the operation triggered by explicit negation. In line with this interpretation are the findings from Rapp and Kendeou (2007) showing that refutation is more likely to occur when people are explicitly instructed to revise their existing knowledge. In this sense, explicit negation might function as an instruction to update knowledge and encode the valid concept. This reasoning is consistent with the Knowledge Revision Components (KReC) Model (Kendeou & O'Brien, 2014) that suggests that knowledge might be revised and updated when a reason (e.g., causal explanation) is provided

why mental correction might be necessary (see also Rapp & Kendeou, 2009). Accordingly, without any reasons to refute false information, false concepts may be more strongly encoded than valid concepts.

A complementary conceptualization attributes the difference between negated and false sentences to the kind of mental model that falsehood implies. Specifically, unlike negation, false information might not necessarily involve the construction of a mental model of a valid concept even when falsehood is detected (see Isberner & Richter, 2013, 2014; Reder, 1982; Richter, 2015; Richter et al., 2009). To illustrate, the sentence "soap is edible" can be rejected on the basis of general world knowledge (e.g., that soap does *not* belong to the category of food) rather than by activating a specific valid alternative (see also Hagoort, Hald, Bastiaansen, & Petersson, 2004; Kutas & Hillyard, 1980). This might be the case even for sentences with a clear valid alternative available. For example, "a pear is a vegetable" might be rejected on the basis of knowledge that a pear does not belong to the category of vegetables. Accordingly, such rejections would not lead to an encoding of valid concepts. It is important to note that the pragmatic implications attributed to negation and the mental model implied by falsehood are not contradictory, and each may contribute to the tendency to prioritize the false concept and neglect the valid concept when being confronted with a false assertion.

We noted earlier that the pragmatic implications of falsehood might require recipients to keep it in mind because such memory can be beneficial in the future. That is, it pays to remember false information in the present, either because it facilitates the construction of counter-arguments that can be used in the future (e.g., the Inoculation Model; McGuire, 1964) or because it helps to identify untrustworthy sources of information so that their messages can be later discounted (Tormala & Clarkson, 2008). Yet, it could be argued that a focus on the valid concept serves as a better means of self-protection. However, to counter the salience of a presented false concept, one has to activate a strong cognitive procedure that allows

inhibition of the salient content. Explicit negation seems to serve this function. Moreover, perhaps because false information was detected so easily, at least in the present paradigm, it created a false sense of security that one is protected from this misinformation. The latter, we believe, is common in real-world cases of bias attributed to identified misinformation.

It is noteworthy that our participants were not instructed to detect or report the truth-value of the sentences. Hence, the present findings might have strong implications for situations in which people detect that messages are false but are not prompted to give explicit truth-value judgments (as may happen in many everyday situations when false information is encountered). The results of all our impression tasks demonstrate the sensitivity of the mental system to truth and falsehood. That is, our respondents reacted positively to true sentences (that happen to be negations of a falsehood) and disliked sentences that conveyed false information (see Unkelbach et al., 2011). Speculatively, the detection of false information might make people worry about being duped (Vohs, Baumeister, & Chin, 2007) or might elicit other negative feelings associated with the discovery of falsehood. The opposite might be the case when one reads a negation of falsehood. Future research should address the question of whether explicit truth-value judgments lead to a stronger encoding of valid concepts and weaker encoding of false concepts when confronted with clearly false information in comparison to explicit negations of falsehood.

Reliance on Misinformation

A typical finding in a paradigm that demonstrates reliance on misinformation (e.g., Fazio et al., 2013) is that even people who reveal factual knowledge in an early test (e.g., "The largest ocean is the Pacific") err in answering questions about this knowledge (e.g., "What is the largest ocean?") after being exposed to misinformation (e.g., "The largest ocean is the Atlantic"). The theoretical analysis of our findings attributes this effect to a difference between encoding of the valid and false concepts while processing misinformation. Accordingly, reliance on misinformation can occur either because false information was

presented in a context that does not support encoding of the valid concept, or because it was presented in a context that supports encoding of the false concept (see Rapp et al., 2014). For example, the context of reading a narrative in which one may be immersed with no goal of truth judgment may lead to both, resulting in an adoption of the false concept even when one has the knowledge to reject it (e.g., Appel & Richter, 2007; Gerrig & Prentice, 1991; Green, 2004; Green & Brock, 2000).

Note that (mis)information may vary with respect to the availability of valid concepts. The sentences in our study utilize false information for which virtually every recipient could come up with the valid concept corresponding to the intended meaning. Of course, in some cases, only one concept is available. For example, the sentence "The MMR vaccine does not lead to autism" (Horne, Powell, Hummel, & Holyoak, 2015; Lewandowsky et al., 2012) attempts to battle a shared misapprehension, namely, that MMR vaccination is responsible for autism in children. However, the general public often lacks the relevant knowledge to clearly grasp a corresponding valid concept. In such cases people are particularly sensitive to misinformation, and attempts to remedy this by using negations are unlikely to succeed (see also Chan et al., 2017). Therefore, it is recommended that remedial attempts should not only employ explicit negations but be accompanied by possible valid representations, such as the risk or the prevalence of diseases the MMR vaccine actually reduces (Horne et al., 2015; see also Ecker, O'Reilly, Reid, & Chang, in press).

The failure to access the correct answer is especially important when people learn information that is later found to be invalid. The belief-perseverance effect (see review in Schul & Burnstein, 1998) and the continued-influence effect (Brydges, Gignac, & Ecker, 2018; Swire, Ecker, & Lewandowsky, 2017) both suggest that after known information has been discredited, recipients do not necessarily access the correct information, especially when their cognitive abilities are taxed. To obtain appropriate correction it is imperative to provide recipients with the alternative (i.e., correct) knowledge so that the original (false) knowledge

can be replaced. In our terminology, it is important to shift them from a stronger encoding of the false concept to a stronger encoding of the valid concept. Ecker et al. (2015) showed that recency plays an important role in the continued influence of misinformation, with more recently presented false concepts showing the strongest influence of misinformation and more recently presented valid concepts showing the strongest correction effects. Thus, correction effects might be most efficient when valid concepts are encoded *during* or even after the processing of false information rather than before false information is processed (e.g., Mazursky & Schul, 1988; see also Rapp, 2008). A similar claim was put forward by Rapp, Hinze, Kohlhepp, and Ryskin (2014). The authors showed that when participants were instructed to correct false information during reading, the influence of false information was substantially reduced.

The proneness to encode the valid information given a communication of misinformation may also depend on the person's state of mind. Past work on distrust (Mayo, 2015; Schul et al., 2004) highlighted that persons who distrust tend to activate cognitions that are incongruent with the given information. Speculatively, the detection of falsehood would be superior by people who distrust; and once misinformation is detected, a mindset of distrust might lead to more resistance against the influence of falsehoods. Future research should address other possible contexts that might foster the activation and encoding of valid concepts. For instance, pragmatic informativeness and recipients' expectations might be relevant factors not only for the processing of negation (Dale & Duran, 2011; Nieuwland & Kuperberg, 2008), they might also determine the encoding of false and valid concepts when confronted with false information. One possible case could be irony (Giora et al., 2005). Investigation of these possibilities should lead to a more comprehensive picture of the influence of context on the processing of false information.

Concluding Remarks

Our study relies on offline memory measures to investigate the encoding of valid and false concepts. At first glance, online measures might be preferable (see Deutsch et al., 2009; Richter et al., 2009). These measures directly assess levels of activation. However, while online measures might be more sensitive in capturing momentary activation of concepts (e.g., Dudschig & Kaup, 2018), we were genuinely interested in the longer-term effects of the differences in encoding, as such effects might be closer to effects of interest, like the reliance on misinformation. Nevertheless, future research should employ both online and offline measures to investigate how the relative activation of true and false concepts influences encoding strength.

Our findings imply that even when false information is detected, it is not mentally corrected to the same degree as false information that is explicitly rejected by negation. Admittedly, the obtained effects in this research are small, referring to behavioral tendencies rather than all-or-none processes. However, these tendencies might be especially influential when people act under uncertainty (e.g., trying to remember whether a piece of information was encountered).

Taken together, our theoretical analysis suggests that false information and the explicit negation of false information are capable of activating both the valid concept and the false concept. However, the encoding strength of the two types of concepts might depend on the context and one's goals at the time the information is processed. At times, people might be extremely proficient and correct false information easily. At other times, the correction of false information might be difficult or even unsuccessful in the sense that people maintain the false concept. It is critical to clarify factors that strengthen the meaning associated with the false and the valid concept in order to predict outcomes when people are confronted with

false information under various conditions. We believe that applying our framework will strongly support this endeavor.

Context

We live in an era of misinformation, which is the palpable context for this study. The explosion of fake news highlights the challenge for people in distinguishing between truth and falsehood, and it shows that they are easily influenced by misinformation. The presented research suggests a theoretical framework that helps to clarify under which conditions people hold fast to falsehoods and which conditions lead to a correction of falsehoods. The main implication of our current findings is that in order to correct false information, it is not enough to trust that people "know better" when they are exposed to obviously false information. Rather, it is critical to explicitly negate falsehood and thereby make recipients consider the valid answer. While even obviously false sentences can bias people away from the truth, our present study suggests that a remedy for this bias is an explicit negation of falsehood, as this seems to trigger consideration of a valid alternative.

References

Adriaanse, M. A., van Oosten, J. M., de Ridder, D. T., de Wit, J. B., & Evers, C. (2011). Planning what not to eat: Ironic effects of implementation intentions negating unhealthy habits. *Personality and Social Psychology Bulletin*, *37*(1), 69–81. doi.org/10.1177/0146167210390523

Anderson, S., Huette, S., Matlock, T., & Spivey, M. J. (2010). On the temporal dynamics of negated perceptual simulations. In F. Parrill, V. Tobin, & M. Turner (Eds.), *Mind, form, and body* (pp. 1–20). Stanford, CA: CSLI Publications.

Appel, M., & Richter, T. (2007). Persuasive effects of fictional narratives increase over time. *Media Psychology, 10*, 113–134. DOI: 10.1080/15213260701301194

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258. dx.doi.org/10.1037/h0055756

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. doi.org/10.1016/j.jml.2012.11.001

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

Beltrán, D., Orenes, I., & Santamaría, C. (2008). Context effects on the spontaneous production of negation. *Intercultural Pragmatics*, *5*(4), 409–419. doi.org/10.1515/IPRG.2008.020

Braasch, J. L. G., Rouet, J. F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in text comprehension. *Memory & Cognition*, *40*, 450–465. doi.org/10.3758/s13421-011-0160-6

Brinol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(1), 49–96. doi.org/10.1080/10463280802643640

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge: University Press.

Brydges, C., Gignac, G., & Ecker, U. (2018). Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. *Intelligence*, *69*, 117–122. doi.org/10.1016/j.intell.2018.03.009

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi.org/10.1177/1745691610393980

Burnstein, E., & Schul, Y. (1982). The informational basis of social judgments: Operations in forming an impression of another person. *Journal of Experimental Social Psychology*, *18*(3), 217–234. doi.org/10.1016/0022-1031(82)90051-8

Burnstein, E., & Schul, Y. (1983). The informational basis of social judgments: Memory for integrated and nonintegrated trait descriptions. *Journal of Experimental Social Psychology*, *19*(1), 49–57. doi.org/10.1016/0022-1031(83)90004-5

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*(1), 45–73. dx.doi.org/10.1037/h0076248

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. doi.org/10.1016/j.chb.2013.05.009

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude

judgment. *Journal of Personality and Social Psychology*, *66*, 460–473. dx.doi.org/10.1037/0022-3514.66.3.460

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science, 28,* 1531–1546. doi.org/10.1177/0956797617714579

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*(3), 472–517. doi.org/10.1016/0010-0285(72)90019-9

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics.* San Diego: Harcourt Brace Jovanovich. doi.org/10.3758/BF03214545

Colston, H. L. (1999). "Not good" is "bad," but "not bad" is not "good": An analysis of three accounts of negation asymmetry. *Discourse Processes*, *28*(3), 237–256. doi.org/10.1080/01638539909545083

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. doi:10.1002/pra2.2015.145052010082

Cook, A. E., & O'Brien, E. J. (2014). Knowledge activation, integration, and validation during narrative text comprehension. *Discourse Processes*, *51*(1-2), 26–49. doi.org/10.1080/0163853X.2013.855107

Craik, F. I. (2002). Levels of processing: Past, present...and future? *Memory*, *10*(5-6), 305–318. doi.org/10.1080/09658210244000135

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. dx.doi.org/10.1037/0096-3445.104.3.268

Cunningham, W. (2015). *Not Cunningham's law*. Retrieved 14 August 2018, from https://www.youtube.com/watch?v=fclyQt6R5Dc.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science, 35*(5), 983–996. doi:10.1111/j.15516709.2010.01164.x

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology, 91*(3), 385–405. doi.org/10.1016/j.jesp.2006.12.004

Deutsch, R., Kordts-Freudinger, R., Gawronski, B., & Strack, F. (2009). Fast and fragile: A new look at the automaticity of negation processing. *Experimental Psychology*, *56*(6), 434–446. dx.doi.org/10.1027/1618-3169.56.6.434

de Vega, M., Morera, Y., León, I., Beltrán, D., Casado, P., & Martín-Loeches, M. (2016). Sentential negation might share neurophysiological mechanisms with action inhibition: Evidence from frontal theta rhythm. *The Journal of Neuroscience*, *36*(22), 6002–6010. doi.org/10.1523/JNEUROSCI.3736-15.2016

Dudschig, C., & Kaup, B. (2018). How does "not left" become "right"? Electrophysiological evidence for a dynamic conflict-bound negation processing account. *Journal of Experimental Psychology: Human Perception and Performance, 44*(5), 716–728. doi.org/10.1037/xhp0000481

Ecker, U. K., O'Reilly, Z., Reid, J. S., & Chang, E. P. (in press). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*. doi.org/10.1111/bjop.12383

Ecker, U. K., Lewandowsky, S., Cheung, C. S., & Maybery, M. T. (2015). He did it! She did it! No, she did not! Multiple causal explanations and the continued influence of misinformation. *Journal of Memory and Language, 85*, 101–115. doi.org/10.1016/j.jml.2015.09.002

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. doi.org/10.3758/BF03193146

Fazio, L. K., Barber, S. J., Rajaram, S., Ornstein, P. A., & Marsh, E. J. (2013). Creating illusions of knowledge: Learning errors that contradict prior knowledge. *Journal of Experimental Psychology: General, 142*, 1–5. dx.doi.org/10.1037/a0028649

Fiedler, K., Walther, E., Armbruster, T., Fay, D., & Naumann, U. (1996). Do you really know what you have seen? Intrusion errors and presupposition effects on constructive memory. *Journal of Experimental Social Psychology*, *32*(5), 484–511. doi.org/10.1006/jesp.1996.0022

Fillenbaum, S. (1966). Memory for gist: Some relevant variables. *Language and Speech*, *9*(4), 217–227. doi.org/10.1177/002383096600900403

Fraenkel, T., & Schul, Y. (2008). The meaning of negated adjectives. *Intercultural Pragmatics*, *5*(4), 517–540. doi.org/10.1515/IPRG.2008.025

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, *44*(2), 370–377. doi.org/10.1016/j.jesp.2006.12.004

Gerrig, R. J., & Prentice, D. A. (1991). The representation of fictional information. *Psychological Science, 2*, 336–340. doi.org/10.1111%2Fj.1467-9280.1991.tb00162.x

Giora, R. (2006). Anything negatives can do affirmatives can do just as well, except for some metaphors. *Journal of Pragmatics*, *38*(7), 981–1014. doi.org/10.1016/j.pragma.2005.12.006

Giora, R., Balaban, N., Fein O., & Alkabetz, I. (2005). Negation as positivity in disguise. In H. L. Colston & A. N. Katz (Eds.), *Figurative language comprehension: Social and cultural influences* (pp. 233–258). Hillsdale, NJ: Erlbaum.

Givón, T. (1978). Negation in language: Pragmatics, function, ontology. *Syntax and Semantics*, *9*, 69–112. doi.org/10.2307/326399

Givón, T. (1993). *English grammar: A function-based introduction* (Vol. 2). Amsterdam: John Benjamins Publishing Company. doi: 10.1075/z.engram1

Grant, S. J., Malaviya, P., & Sternthal, B. (2004). The influence of negation on product evaluations. *Journal of Consumer Research, 31*(3), 583–591. doi.org/10.1086/425093

Green, M. C. (2004). Transportation into narrative worlds: The role of prior knowledge and perceived realism. *Discourse Processes, 38*, 247–266. doi.org/10.1207/s15326950dp3802_5

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79*, 701-721. psycnet.apa.org/doi/10.1037/0022-3514.79.5.701

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3, Speech acts* (pp. 43–58). New York: Academic Press.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438–441. doi.org/10.1126/science.1095455

Haran, D., Mor, N., & Mayo, R. (2011). Negating in order to be negative: The relationship between depressive rumination, message content and negation processing. *Emotion, 11*(5), 1105–1111. dx.doi.org/10.1037/a0025301

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation? An examination of negated metaphors. *Journal of Pragmatics, 38*(7), 1015-1032. doi.org/10.1016/j.pragma.2005.12.005

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review, 51*(6), 358–374. dx.doi.org/10.1037/h0055425

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*(2), 305–324. doi.org/10.3758/s13421-013-0359-9

Horn, L. R. (1989). *A natural history of negation.* Chicago: University of Chicago Press.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*(33), 10321–10324. doi.org/10.1073/pnas.1504019112

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635–650. doi.org/10.1086/266350

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2018). When people co-occur with good or bad events: Graded effects of relational qualifiers on evaluative conditioning. *Personality and Social Psychology Bulletin*, *45*(2), 196–208. doi.org/10.1177/0146167218781340

Hunt, E. (2016). What is fake news? How to spot it and what you can do to stop it. *The Guardian*. Retrieved 17 March 2019, from https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate.

Inquisit 4.0.6.0 [Computer software]. (2014). Seattle, WA: Millisecond Software.

Isberner, M. B., & Richter, T. (2013). Can readers ignore implausibility? Evidence for nonstrategic monitoring of event-based plausibility in language comprehension. *Acta Psychologica*, *142*(1), 15–22. doi.org/10.1016/j.actpsy.2012.10.003

Isberner, M. B., & Richter, T. (2014). Does validation during language comprehension depend on an evaluative mindset? *Discourse Processes*, *51*(1-2), 7–25. doi.org/10.1080/0163853X.2013.855867

Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, *71*(3), 191–229. doi.org/10.1016/S0010-0277(99)00015-3

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. doi.org/10.1016/0010-0285(76)90015-3

Kaup, B. (2001). Negation and its impact on the accessibility of text information. *Memory & Cognition*, *29*(7), 960–967. doi.org/10.3758/BF03195758

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*(7), 1033–1050. doi.org/10.1016/j.pragma.2005.09.012

Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 439–446. dx.doi.org/10.1037/0278-7393.29.3.439

Kaup, B., Zwaan, R. A., & Lüdtke J. (2007). The experiential view of language comprehension: How is negated text information represented? In Schmalhofer, F. & Perfetti C. A. (Eds.), *Higher-level language processes in the brain: Inference and comprehension processes* (pp. 255–288). Mahwah, NJ: Erlbaum.

Kendeou, P., & O'Brien, E. J. (2014). The knowledge revision components (KReC) framework: Processes and mechanisms. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 353–377). Cambridge, MA: MIT Press.

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

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. doi.org/10.1126/science.7350657

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. doi.org/10.1126/science.aao2998

Lea, R. B., & Mulligan, E. J. (2002). The effect of negation on deductive inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 303–317. dx.doi.org/10.1037/0278-7393.28.2.303

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. doi.org/10.1177/1529100612451018

MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 633–642. dx.doi.org/10.1037/0278-7393.15.4.633

Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, *26*(1), 283–327. doi.org/10.1080/10463283.2015.1117249

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs. "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433–449. doi.org/10.1016/j.jesp.2003.07.008

Mayo, R., Schul, Y., & Rosenthal, M. (2014). If you negate, you may forget: Negated repetitions impair memory compared with affirmative repetitions. *Journal of Experimental Psychology: General*, *143*(4), 1541–1552. dx.doi.org/10.1037/a0036122

Mazursky, D., & Schul, Y. (1988). The effects of advertisement encoding on the failure to discount information: Implications for the sleeper effect. *Journal of Consumer Research*, *15*, 24-36. doi.org/10.1086/209142

McGeady, S. (2010). Schott's Vocab, *The New York Times*. Retrieved 14 August 2018, from https://schott.blogs.nytimes.com/2010/05/31/jurisimprudence/.

McGuire, W. J. (1964). Inducing resistance to persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1, pp. 191–229). New York, NY: Academic Press.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*(2-3), 131–157. doi.org/10.1080/01638539809545042

Nadarevic, L., & Erdfelder, E. (2013). Spinoza's error: Memory for truth and falsity. *Memory & Cognition*, *41*(2), 176–186. doi.org/10.3758/s13421-012-0251-z

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218. doi.org/10.1111/j.1467-9280.2008.02226.x

Nordmeyer, A. E., & Frank, M. C. (2014). A pragmatic account of the processing of negative sentences. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. escholarship.org/uc/item/95j3b18w

Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language, 74*, 36–45. doi.org/10.1016/j.jml.2014.04.001

Pantazi, M., Kissine, M., & Klein, O. (2018). The power of the truth bias: False information affects memory and judgment even in the absence of distraction. *Social Cognition*, *36*(2), 167–198. doi.org/10.1521/soco.2018.36.2.167

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. doi.org/10.1016/j.jesp.2017.01.006

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. doi.org/10.1146/annurev.psych.54.101601.145041

Phillips, P. J., Moon, H., Rizvi, S.A., & Rauss P. J. (2000). The FERET Evaluation Methodology for Face Recognition Algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 1090-1104. doi.org/10.1109/34.879790

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, *16*(5), 295-306. doi.org/10.1016/S0262-8856(97)00070-X

Rapp, D. N. (2008). How do readers handle incorrect information during reading? *Memory & Cognition*, *36*(3), 688–701.

Rapp, D. N. (2016). The consequences of reading inaccurate information. *Current Directions in Psychological Science*, *25*(4), 281–285. doi.org/10.1177/0963721416649347

Rapp, D. N., & Braasch, J. L. (2014). *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. Cambridge, MA: MIT Press.

Rapp, D. N., Hinze, S. R., Kohlhepp, K., & Ryskin, R. A. (2014). Reducing reliance on inaccurate information. *Memory & Cognition*, *42*(1), 11-26. https://doi.org/10.3758/s13421-013-0339-0

Rapp, D. N., Hinze, S. R., Slaten, D. G., & Horton, W. S. (2014). Amazing stories: Acquiring and avoiding inaccurate information from fiction. *Discourse Processes*, *51*(1-2), 50–74. doi.org/10.1080/0163853X.2013.855048

Rapp, D. N., & Kendeou, P. (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition*, *35*(8), 2019-2032. doi.org/10.3758/bf03192934

Rapp, D. N., & Kendeou, P. (2009). Noticing and revising discrepancies as texts unfold. *Discourse Processes*, *46*(1), 1–24. doi.org/10.1080/01638530802629141

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, *89*(3), 250–280. doi.org/10.1037//0033-295x.89.3.250

Richter, T. (2015). Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes*, *52*(5-6), 337–355. doi.org/10.1080/0163853X.2015.1025665

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*(3), 538–558. dx.doi.org/10.1037/a0014038

Sachs, J. S. (1974). Memory in reading and listening to discourse. *Memory & Cognition*, *2*(1), 95–100. doi.org/10.3758/BF03197498

Schul, Y., & Burnstein, E. (1998). Suspicion and discounting: Ignoring invalid information in an uncertain environment. In Golding, J. M. & MacLeod, C. M. (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 321–348). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers. doi.org/10.1080/0163853X.2013.855534

Schul, Y., & Mayo, R. (2014). Discounting information: When false information is preserved and when it is not. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 203–220). Cambridge, MA: MIT Press.

Schul, Y., Mayo, R., & Burnstein, E. (2004). Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *Journal of Personality and Social Psychology*, *86*(5), 668–679. dx.doi.org/10.1037/0022-3514.86.5.668

Schul, Y., & Mazursky, D. (1990). Conditions facilitating successful discounting in consumer decision making. *Journal of Consumer Research*, *16*(4), 442–451. https://doi.org/10.1086/209229

Singer, M. (2013). Validation in reading comprehension. *Current Directions in Psychological Science*, *22*(5), 361–366. doi.org/10.1177/0963721413495236

Singer, M. (2019). Challenges in Processes of Validation and Comprehension. *Discourse Processes*, 1-19. https://doi.org/10.1080/0163853X.2019.1598167

Singer, M., & Doering, J. C. (2014). Exploring individual differences in language validation. *Discourse Processes*, *51*(1-2), 167–188. doi.org/10.1080/0163853X.2013.855534

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin*, *39*(2), 193–205. doi.org/10.1177/0146167212472374

Sparks, J. R., & Rapp, D. N. (2011). Readers' reliance on source credibility in the service of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 230–247. doi.org/10.1037/a0021331

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(12), 1948–1961. doi.org/10.1037/xlm0000422

Tian, Y., Ferguson, H., & Breheny, R. (2016). Processing negation without context: Why and when we represent the positive argument. *Language, Cognition and Neuroscience*, *31*(5), 683–698. doi.org/10.1080/23273798.2016.1140214

Tormala, Z. L., & Clarkson, J. J. (2008). Source trustworthiness and information processing in multiple message situations: A contextual analysis. *Social Cognition, 26*, 357–367. doi.org/10.1521/soco.2008.26.3.357

Unkelbach, C., Bayer, M., Alves, H., Koch, A., & Stahl, C. (2011). Fluency and positivity as possible causes of the truth effect. *Consciousness and Cognition*, *20*(3), 594–602. doi.org/10.1016/j.concog.2010.09.015

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Vohs, K. D., Baumeister, R. F., & Chin, J. (2007). Feeling duped: Emotional, motivational, and cognitive aspects of being exploited by others. *Review of General Psychology, 11*(2), 127–141. dx.doi.org/10.1037/1089-2680.11.2.127

Wyer, R. S., & Srull, T. K. (1986). Human cognition in its social context. *Psychological Review*, *93*(3), 322–359. hdl.handle.net/1783.1/47227

Xu, A. J., & Wyer, R. S. (2012). The role of bolstering and counterarguing mind-sets in persuasion. *Journal of Consumer Research*, *38*(5), 920–932. doi.org/10.1086/661112

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129–166. doi/10.1007/BF02249397

APPENDIX

Overview of experimental sentences

Affirmative false	Negated true	Baseline
The sun rises in the west.	The sun does not rise in the west.	The sun rises after dawn.
Turtles move fast.	Turtles do not move fast.	Turtles eat lettuce.
Giraffes are short.	Giraffes are not short.	Giraffes inhabit savannas.
A forest consists of dunes.	A forest does not consist of dunes.	A forest is a habitat.
Scissors are used to paste.	Scissors are not used to paste.	Scissors are used in school.
England is a city.	England is not a city.	England has a soccer team.
Boats sail on land.	Boats do not sail on land.	Boats have sails.
Shorts are worn in winter.	Shorts are not worn in winter.	Shorts are worn by kids.
Sugar is bitter.	Sugar is not bitter.	Sugar is used for desserts.
Lemons are salty.	Lemons are not salty.	Lemons are juicy.
The Arctic is hot.	The Arctic is not hot.	The Arctic is located in the Arctic Ocean.
Fish have pores.	Fish do not have pores.	Fish have bones.
Birds have fur.	Birds do not have fur.	Birds have wings.
Beer is solid.	Beer is not solid.	Beer contains alcohol.
Soap makes you dirty.	Soap does not make you dirty.	Soap is a hygiene product.
Towels are used to get wet.	Towels are not used to get wet.	Towels can be found in the bathroom.
Cars have legs.	Cars do not have legs.	Cars have an ignition lock.

Omelets are made from chicken.	Omelets are not made from chicken.	Omelets can be made with parsley.
The "Big Apple" is Detroit.	The "Big Apple" is not Detroit.	The "Big Apple" is a nickname.
A dove is a symbol of war.	A dove is not a symbol of war.	A dove is a sign of hope.
You should cross the road when your traffic light is red.	You should not cross the road when your traffic light is red.	You should cross the road when the traffic light signals it.
Lightning flashes are dark.	Lightning flashes are not dark.	Lightning flashes can be seen during a storm.
Elephants are small.	Elephants are not small.	Elephants have a trunk.
Airplanes fly in the streets.	Airplanes do not fly in the streets.	Airplanes fly at high altitudes.
Most brides wear black.	Most brides do not wear black.	Most brides wear dresses.
Rocks are soft.	Rocks are not soft.	Rocks are minerals.
Jumbo jets are light.	Jumbo jets are not light.	Jumbo jets are a means of transport.
A pear is a vegetable.	A pear is not a vegetable.	A pear has seeds inside.
Baked beans is a beverage.	Baked beans is not a beverage.	Baked beans can be bought in cans.
Diamonds are cheap.	Diamonds are not cheap.	Diamonds are gemstones.
Most people sleep during the day.	Most people do not sleep during the day.	Most people sleep in a comfortable position.
Peas are square.	Peas are not square.	Peas are edible raw and cooked.
An hour has 60 seconds.	An hour does not have 60 seconds.	An hour is defined as a period of time.

Magazines are made from textile.	Magazines are not made from textile.	Magazines are sold at newsstands.
You lock a door with a knife.	You do not lock a door with a knife.	You lock a door to prevent others from entering.
During a theatre performance the audience should be noisy.	During a theatre performance the audience should not be noisy.	During a theatre performance the audience is sitting.
The lowest story of a building is the attic.	The lowest story of a building is not the attic.	The lowest story of a building is sometimes underground.
Cake is baked in the fridge.	Cake is not baked in the fridge.	Cake is baked on special occasions.
Fever, coughing, and a running nose are signs that you are healthy.	Fever, coughing, and a running nose are signs that you are not healthy.	Fever, coughing, and a running nose are symptoms of a flu.
Winning the lottery is common.	Winning the lottery is not common.	Winning the lottery is a stroke of luck.
Trains run on highways.	Trains do not run on highways.	Trains stop at railroad stations.
Honey is made by butterflies.	Honey is not made by butterflies.	Honey is made out of nectar.
Most people’s dominant hand is left.	Most people’s dominant hand is not left.	Most people’s dominant hand is more skillful.
You need to charge your battery when it is full.	You need to charge your battery when it is not full.	You need to charge your battery once in a while.
Texas lies in the north.	Texas does not lie in the north.	Texas is in the USA.

A chair is a piece of clothing.	A chair is not a piece of clothing.	A chair is an object.
Alligators are mammals.	Alligators are not mammals.	Alligators are about 13 ft long.
An eye of a needle is wide.	An eye of a needle is not wide.	An eye of a needle is made for pulling thread.
The currency in the USA is dinar.	The currency in the USA is not dinar.	The currency in the USA consists of bills and coins.
Dogs wag their whiskers.	Dogs do not wag their whiskers.	Dogs wag when they are joyful.
People smile when they are angry.	People smile when they are not angry.	People smile when something funny happens.
Breakfast is eaten in the evening.	Breakfast is not eaten in the evening.	Breakfast is eaten all over the world.
Soup is eaten with a fork.	Soup is not eaten with a fork.	Soup is eaten as a starter.
You leave the subway through the entrance.	You do not leave the subway through the entrance.	You leave the subway when you have reached your destination.
People lie down to sleep in the kitchen.	People do not lie down to sleep in the kitchen.	People lie down to sleep when they are exhausted.
You chew with your tongue.	You do not chew with your tongue.	You chew with your jaw.
People drink coffee from a pan.	People do not drink coffee from a pan.	People drink coffee at work.
Scarves are worn around the wrist.	Scarves are not worn around the wrist.	Scarves are worn in the fall.
Shoes are worn on hands.	Shoes are not worn on hands.	Shoes are worn by humans.

Belts are worn around the ankle.	Belts are not worn around the ankle.	Belts are worn to support pants.

*Presentation and counterbalancing: Sentences appeared in a false affirmative, a true negated, or a baseline version, as a between-participants manipulation, and were never repeated within-participant. Thus, each participant saw as experimental trials 20 different affirmative false, 20 different negated true, and 20 different baseline sentences. Which sentence appeared as affirmative false, negated true, or baseline version was counterbalanced between-participants. To this end, we created six different counterbalancing conditions. The counterbalancing ensured that every sentence appeared with the same frequency as affirmative false, negated true, or baseline version, and it guaranteed an independence of version and nominal sentence. Sentences appeared in a different random order for each participant.