LUNDS UNIVERSITET

DEPARTMENT of PSYCHOLOGY

Avoiding Misattribution Errors through Recollection: Can Retrieval Practice Suppress the Illusory Truth Effect?

Tim Brabender

Master's Thesis (30 hp)
Spring 2023

Supervisors: Mikael Johansson, Inês Bramão

Acknowledgements

I would like to thank everyone that supported me throughout the process of creating this work. Especially, I would like to thank my supervisors, Mikael Johansson and Inês Bramão, for their ongoing work in helping me design this study and guiding me towards the finished paper.

I would also like to thank my fellow master's student Hannah Tucker, as her unique perspective on psychological processes inspired me towards implementing mixed-effects modelling in my analysis.

Finally, I would like to thank my family and friends in providing a firm base of support to rely on throughout this project.

Abstract

Misinformation is a major global issue and presents major negative impacts on many aspects of our lives. One of the factors facilitating misinformation belief is the illusory truth effect, the phenomenon of how repeated exposure to information can increase subjective truth. This occurs when processing fluency, caused by repeated exposure to an item, is misattributed to the truth status of an item. This study investigates if retrieval practice can suppress the illusory truth effect, after the first exposure to an information already happened. In a within-person study, participants were exposed to several trivia statements during an initial exposure phase. In a re-exposure phase, retrieval practice and a non-retrieval task were presented in alternating blocks. Finally, there was a truth rating phase accompanied by a warning about the illusory truth effect. Results indicate that there was a very small illusory truth effect. The retrieval practice intervention did not suppress the illusory truth effect when compared to the non-retrieval task. Furthermore, examining the effect of processing fluency revealed that successfully recollected statements were processed more fluently than when they were not recollected. The findings suggest that recollection alone cannot suppress the illusory truth effect and that more emphasis should be placed on the process of correctly attributing processing fluency to a prior exposure. This study contributes to the discourse about effective measures to reduce the illusory truth effect and misinformation belief.

Keywords: illusory truth effect, misinformation, retrieval practice, recollection, misattribution, processing fluency

Avoiding Misattribution Errors through Recollection: Can Retrieval Practice Suppress the Illusory Truth Effect?

The distribution and belief in false information is a major global issue, with increasing negative effects on public health (van der Linden, 2022; Zarocostas, 2020), climate change (Treen et al., 2020) and democracy (Lewandowsky et al., 2017). Recent events such as the COVID-19 pandemic as well as the 2020 US elections were accompanied by a growing research interest surrounding phenomena like misinformation and disinformation (Pérez-Escolar et al., 2023). Going forward, it is therefore crucial to investigate why people believe in information that is factually false and how we can counter that (Vosoughi et al., 2018).

False information is now more present than ever in our daily lives, with its spread being especially enabled through social media (Vosoughi et al., 2018). However, we need to first clarify what kinds of false information there are and what are being discussed in this study. Misinformation is mainly distinguishable from e.g., disinformation by the intent of the person supplying it, where disinformation is used with an intention to deceive, whereas misinformation is not (Aïmeur et al., 2023; Allen et al., 2020). Another popular iteration of false information is Fake news. These are falsehoods in the format of news headlines that want to give the intention of a real event that happened (Lazer et al., 2018). All of the above are commonly distributed on social media either with malicious intent or due to people not examining information for truth beforehand (Pennycook & Rand, 2021). The focus of this study will be on misinformation.

Illusory Truth Effect

One major contributor to the belief in misinformation is the illusory truth effect (Ecker et al., 2022). The illusory truth effect refers to the phenomenon of how repeating unfamiliar claims increases subjective truth (Dechêne et al., 2010). In their seminal study, Hasher et al. (1977) demonstrated that when exposing participants to statements at different occasions, as well as repeating some statements, people tended to rate repeated statements as more truthful than non-repeated ones. This effect has been replicated across many domains, using news headlines (Pennycook

et al., 2018), varying time periods (Garcia-Marques et al., 2016) and even opinion statements (Arkes et al., 1989).

The size of the illusory truth effect is moderated by several factors. Dechêne et al. (2010) investigated various moderator variables in a meta-analysis, including different truth measurements, presentation styles, data collection types and levels of processing during the first encounter. For example, the illusory truth effect was smaller when studies used an uneven scale, compared to forcing participants into a dichotomous true-false decision (Dechêne et al., 2010).

Furthermore, there seems to be a relationship between levels of processing, meaning different levels of cognitive involvement, at the initial exposure, and the size of the illusory truth effect (Dechêne et al., 2010). Hawkins and Hoch (1992) found a smaller illusory truth effect when people rate the truthfulness of a stimuli at first exposure, which requires high cognitive involvement, compared to when people rate the comprehensibility, which only requires low cognitive involvement. The relationship between levels of processing and the magnitude of the illusory truth effect is controversial, however, as Unkelbach and Rom (2017) found a that self-referential thinking, one of the higher levels of processing (Craik & Lockhart, 1972), actually enhances the illusory truth effect.

Processing Fluency

One of the mechanisms enabling the illusory truth effect is processing fluency (Alter & Oppenheimer, 2009; Dechêne et al., 2010; Unkelbach et al., 2019). This phenomenon describes the subjective experience that results from stimuli that are easier, or more fluent, to process, opposed to stimuli that are proportionately harder, or more disfluent, to process (Alter & Oppenheimer, 2009). As Jacoby and Dallas (1981) demonstrated, repeatedly encountering a certain information can lead to that information being easier to process later on, therefore enhancing fluency. Reber and Schwarz (1999) linked that experience of processing fluency to an increase in subjective truth. The manipulation of processing fluency in their study was achieved through changing the color contrast with which stimuli statements, in this case geography-related claims, were presented. The results indicated that stimuli which

were processed more fluently were rated subjectively more true (Reber & Schwarz, 1999). The same finding was replicated by Hansen et al. (2008) using consumer-related statements. However, it was added that the effect on subjective truth only persists when a relative change in fluency is detected. This means that the effect was only present when a statements was perceived relatively more fluent, or disfluent, than the former (Hansen et al., 2008).

Unkelbach and Rom (2017) offer an explanation on how repetition might cause processing fluency, suggesting that truth judgements are based on references in our memory that are coherent with the information at hand. Coherence is achieved when an information that people are processing matches references in memory, and according to Unkelbach and Rom (2017), coherence results in more fluent processing, and therefore the illusory truth effect.

Misattribution

Processing fluency drives the illusory truth effect as people learn to use the feeling of fluency as a cue for their truth judgements, especially, when there is no available knowledge that might inform the truth decision (Unkelbach & Greifeneder, 2013). For this to happen, however, people need to attribute the experience of fluency to a task-relevant source, opposed to assuming it is background noise, as otherwise its influence on the present judgement disappears (Alter & Oppenheimer, 2009; Unkelbach & Greifeneder, 2013).

What source it is attributed can change depending on the task. For example, Jacoby et al. (1989) demonstrated that, when participants were asked to rate if names were famous, non-famous names were judged increasingly famous when they were repeated 24 hours later. Similar effects have been found across many other domains, e.g. liking, confidence and frequency (Alter & Oppenheimer, 2009). For truth judgements, the direction of the fluency attribution is naturally one-sided. In a recent study, Corneille et al. (2020) tested whether fluency would still enhance the subjective truth of an information, even when participants are asked to rate if a statement is false or not, thereby reversing the context of the task. When participants were asked to rate whether a statement was false, repeated statements were still

rated more true than new statements. Controversially, when participants were to indicate whether a statement has been previously used as fake news, repeated statements were more likely to be identified as such than new statements (Corneille et al., 2020).

The association of fluency and truth is ecologically valid, as in our day-to-day lives, there usually is a positive correlation between fluency experience and truth (Unkelbach, 2007). Furthermore, hearing a statement multiple times gives the impression of social consensus, which influences perceived truth (Schwarz et al., 2007). However, Unkelbach (2007) showed that the fluency-truth association can be reversed by teaching participants that truth and fluency is negatively correlated. This was done by manipulating fluency with color contrasts and setting participants up in an initial session, where they learn that fluency and truth is either positively or negatively correlated. Results showed that in a later truth rating phase, participants tended to rate more fluent stimuli as false, if they initially learned that fluency and truth are negatively correlated (Unkelbach, 2007).

Another viable technique to counter the effect of processing fluency is fluency discounting. This means that people register the fluency experience, however choose not to use it as a cue in their judgement, either because they decide the feeling is irrelevant to the task at hand, or because they attribute fluency to its correct source (Alter & Oppenheimer, 2009).

Actively Reducing Illusory Truth

Several studies have explored active measures to reduce the illusory truth effect. Nadarevic and Erdfelder (2014) discovered that an initial accuracy focus, meaning the rating of the truth value of a statement at the first exposure, can reduce the illusory truth effect. The same findings have been replicated by Brashier et al. (2020) as well as Calvillo and Smelter (2020), who used news headlines instead of trivia statements. Kruijt et al. (2022) further demonstrate how a small critical thinking recommendation effectively helped people distinguishing true from false messages about Covid-19 on social media.

Nadarevic and Aßfalg (2017) investigated whether warning participants about the illusory truth effect, after the first exposure already happened, could reduce the illusory truth effect (Nadarevic & Aßfalg, 2017). This warning, combined with asking participants to actively avoid it, resulted in a reduction of the illusory truth effect in the truth rating phase, however did not eliminate it (Nadarevic & Aßfalg, 2017). They argue that participants failed to attribute the processing fluency caused by repeated stimuli to an encounter inside the experiment, and instead misattributed it to a previous encounter outside the experiment (Nadarevic & Aßfalg, 2017). This hints at successful recollection of source memory as an important factor in reducing the illusory truth effect.

Familiarity and Recollection

When evaluating influences on the illusory truth effect, two memory processes are especially important, familiarity and recollection. Familiarity refers to the simple process as recognizing something as having encountered before, without knowing when or at what occasion the first encounter was (Yonelinas, 2002). Familiarity processes are often a roadblock for traditional debunking and fact-checking efforts, as both of those strategies usually involve repeating the claim that is to be refuted, which increases its familiarity and therefore might cause people to belief even more in it (Ecker et al., 2020; Swire et al., 2017). The idea that familiarity increases the illusory truth effect is supported by brain imaging evidence suggesting that the illusory truth effect is accompanied by stronger brain activations in the perirhinal cortex (PRC), which is a brain region largely responsible for processing familiarity and retrieval of semantic memory (Eichenbaum et al., 2007; Skinner & Fernandes, 2007; Wang et al., 2016). The functional localization of the illusory truth effect in the PRC also emphasizes the tight relationship between familiarity and processing fluency, where processing fluency is mediating PRC activity, which affects if an item is judged as familiar (Dew & Cabeza, 2013).

Recollection has been identified as a possible countering factor of the illusory truth effect. Recollection adds the component of remembering specific contextual information, or source memory, from the first encounter (Yonelinas, 2002). Begg et al. (1992) showed how successful recollection of source memory can decrease the

magnitude of the illusory truth effect. In the original study, statements were paired with cue words that served as a source. When evaluating the statements later, the illusory truth effect was significantly smaller when the cue words were successfully recollected (Begg et al., 1992).

When it comes to recollection, brain imaging evidence suggests that the encoding process is crucial for an item to be later recollected, as encoding processes for familiarity and recollection are distinct (Sadeh et al., 2012). Mitchell et al. (2005) demonstrated that increased encoding activity in brain areas commonly linked to recollective encoding, the hippocampus and ventrolateral prefrontal cortex (vlPFC), led to a smaller illusory truth effect. They furthermore showed that during decoding, a reduction of the illusory truth effect was linked to stronger brain activations in areas associated with recollective retrieval, namely the inferior frontal gyrus (IFG), the left superior frontal gyrus (SFG) and posterior parietal regions (Mitchell et al., 2005).

The relationship between recollection and processing fluency is complex and depends on whether the fluency occurs during encoding or retrieval of information. Li et al. (2015) demonstrated in an electrophysiological experiment, that when processing fluency is present during the encoding stage, stimuli are less likely to be recollected later. In terms of the illusory truth effect, this could suggest that fluent encoding prevents later recollection and therefore inhibits the effect of successful recollection on the illusory truth effect. During decoding, better recollection of source memory for the first encounter is associated with higher processing fluency (Huang & Shanks, 2021). This is still in line with the notion that recollection suppresses the illusory truth effect, as processing fluency only drives the effect when it is incorrectly attributed to truth, and fluency can be discounted when source memory for the first exposure is recollected (Alter & Oppenheimer, 2009).

Unfortunately, people do not always remember where they heard information first, leading to the question of how to improve recollection. In a recent study, Guran et al. (2020) demonstrated that retrieval practice, the procedure of repeatedly retrieving memory to improve memory, can enhance recollection. In their study setup, participants performed an initial encoding phase, followed by two intermixed task blocks, where one task was a simple categorization task, and one task was an old-

new recognition task. Finally, participants had to undergo a final old-new recognition task. The results show that recollection performance was significantly better for stimuli which were part of the recognition task, suggesting that retrieval practice can enhance the subsequent recollection of a presented item (Guran et al., 2020).

The Present Study

For now, we have established that the illusory truth effect is driven by processing fluency, and countering processing fluency can be achieved through discounting the fluency experience by remembering contextual information from the encoding stage (Alter & Oppenheimer, 2009; Dechêne et al., 2010). This study attempts to equip people with all those prerequisites and investigate whether this proves an effective solution against the illusory truth effect.

In this study, participants were exposed to trivia statements in three phases. For the initial exposure phase, they rated statements for how interesting they find them. In a re-exposure phase, two alternating training tasks were completed. New and old statements were presented, and participants had to perform an old-new recognition task, which served as the retrieval practice, or assign statements to a category, which was the non-retrieval control task. In a final test phase, a warning was presented to the participants, informing them about the illusory truth effect and how it affects truth judgements. Then, participants rated old and new statements for subjective truth.

The aim of this study is to suppress the illusory truth effect by using a unique approach that incorporates recollection into an active intervention, countering the illusory truth effect after the initial exposure already happened.

The results of this study could further inform the creation of effective interventions against the illusory truth effect and shape the way we handle misinformation and fake news. Ultimately, this study could advance the fight against misinformation belief and its negative impact on many aspects of our lives, including health care, climate change and democracy.

Hypotheses

H1: There will be an illusory truth effect, where repeated statements will be rated significantly higher than new statements. Specifically, statements from the first phase should be assigned higher truth ratings than the second phase, and statements from the second phase should be assigned higher truth ratings than the third phase. As Nadarevic and Aßfalg (2017) demonstrate, there was still an illusory truth effect when participants were warned about it, even though it was significantly smaller (d_z = .35). The same should be expected to apply in this study. We also expect that the effect should persist irrespective of the objective truth value of a statement.

H2: Retrieval Practice will have an impact on the suppression of the illusory truth effect. We expect the illusory truth effect to be smaller for statements that were part of the retrieval practice than for statements that were part of the non-retrieval control task. Nadarevic and Aßfalg (2017) argue that the illusory truth effect was not fully eliminated because participants could not attribute processing fluency to the encounter during the experiment, meaning they did not remember encountering it during the experiment. Helping participants to successfully recollect the previous exposure, using retrieval practice as in Guran et al. (2020), during the experiment could facilitate the reduction of the illusory truth effect.

Method

Participants

The sample for this study was sourced online using the social media platform Reddit as well as the platforms SurveySwap and SurveyCircle. All participants were screened for being at least 18 years old, having normal or corrected-to normal vision and having no neurological disorders. In total, 96 people completed the full study. Participants were excluded from the analysis if they did not provide enough answers in the final rating task, where the threshold was set at 70% of the items being answered. Furthermore, to ensure that engagement was given, only participants were included whose performance in the non-retrieval task was above chance level and within two standard deviations (*SD* = .17) from the mean (*M* = .71). The resulting inclusion criteria was a hit rate in the non-retrieval task of at least 38%. Additionally, participants were excluded when their mean reaction time in the final truth rating task

was less than 1000ms, as in Unkelbach (2007). Out of all participants who completed the study, *N* = 88 met all inclusion criteria. The final sample consisted of 36 female, 48 male and 4 non-binary participants. The mean age was *M* = 27.1 (*SD* = 8.92) Participants came from 31 different countries and 4 different continents, including North America, Europe, Asia, and Oceania.

Materials

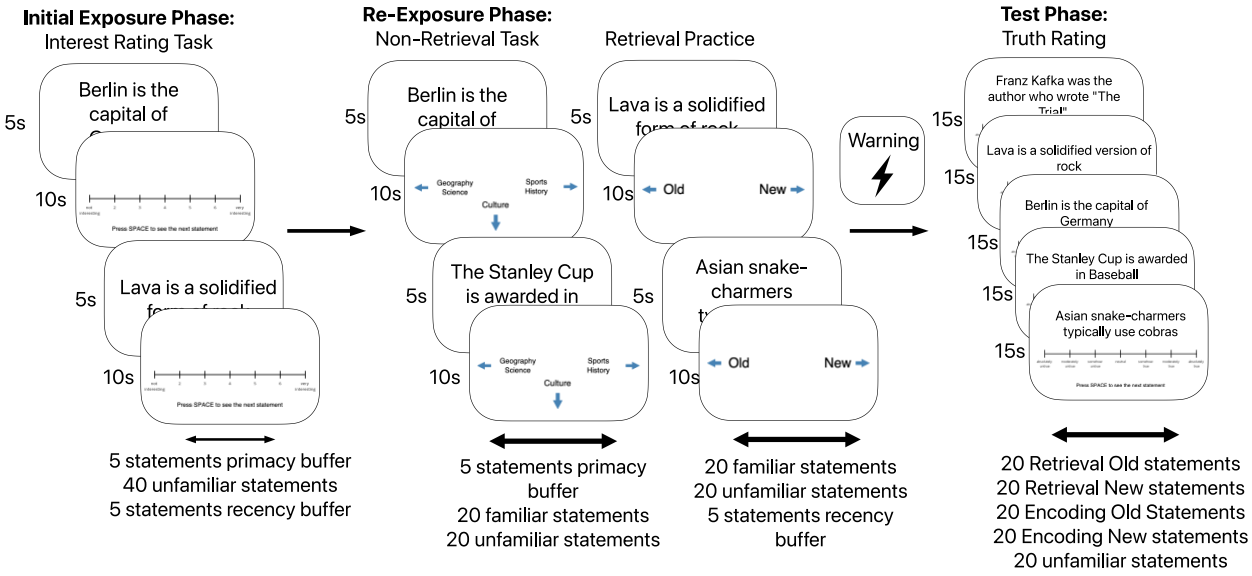
The materials used in this study consisted of 100 normed trivia statements taken from Wertgen and Richter (2023), where the selected statements ranged widely in their recall norming (.14 to .98). The material was converted from a question format to a statement, e.g. “What is the capital of Germany? Berlin” became “Berlin is the capital of Germany”. Furthermore, half of the questions were altered to false statements, e.g. “The Bible is the holy book of the Christians” became “The Bible is the holy book of the Muslims”. Finally, statements were selected to fit a more general sample as the original norming study was done with a German sample. To do this, statements were avoided that contained information very specific to German culture, e.g. “The German Bundestag (Parliament) meets in Berlin”. To make sure that differences across phases and conditions cannot be due to the material, the statements were balanced across five different sets by the mean recall norming provided by Wertgen and Richter (2023), mean statement length, categories, and truth status. Each set contained 20 statements. All sets contained equal amounts of true and false statements, which is similar to previous studies on the illusory truth effect (Calvillo & Smelter, 2020; Nadarevic & Aßfalg, 2017).

Procedure and Design

The study was conducted online and followed a within-subject, blocked design with three phases, an initial exposure, a re-exposure, and a test phase. Participants were briefed using Qualtrics (Qualtrics, 2020), where consent had to be given. Then, participants were redirected to the experiment which was created using PsychoPy (Peirce et al., 2019) and hosted online through Pavlovia.

Figure 1

Experimental Paradigm of the Conducted Study



Note. This figure illustrates the paradigm chosen for this study. The proportions of the items are altered for better readability and do not accurately represent how items appeared during the experiment. In the initial exposure phase, there was an interest rating task where 40 new stimuli were rated for interest. In Phase 2, two alternating tasks were presented, a retrieval practice task and a non-retrieval task. For both tasks, 20 stimuli from phase 1 and 20 new stimuli were presented. Before the final rating test, a warning was presented to the participants about illusory truth. Then, in the test phase, all stimuli from the previous phases and 20 new stimuli were rated for truthfulness.

Figure 1 illustrates the experimental paradigm chosen for this study. Before the study phase, participants were briefed on the general layout of the experiment. Then, for the initial exposure phase, participants rated 40 trivia statements for how interesting they find them. Statements were shown for 5 s after which participants had 10 s to rate statements on a 7-point scale from “uninteresting” to “very interesting”. There was a primacy and recency buffer of 5 statements, resulting in a total of 50 trials for the initial exposure phase.

The re-exposure phase followed immediately, where two different, alternating, blocked tasks were presented. There was a retrieval practice task, where participants were asked to indicate if a statement has been shown in the initial exposure phase or if it is new. Half of these statements were previously shown in the initial exposure phase, and half were new. The non-retrieval task was a categorization task, where participants were asked to assign statements to five different categories, Geography, Science, History, Sports, and Culture. Half of the statements were old statements

previously shown in the initial exposure phase, and half were new. All statements were shown for 5 s after which participants had 10 s to perform the specific task. The different tasks alternated in blocks of ten stimuli and were divided by an instruction to what the following task will be. There were 4 blocks for each task, as well as a primacy and a recency buffer of 5 stimuli each, resulting in a total of 90 trials for the re-exposure phase.

Immediately after the re-exposure phase, the test phase followed. Participants were warned about the illusory truth effect and how it biases truth ratings and were asked to prevent the illusory truth effect in the following task. The warning read as follows:

“However, before this phase begins, you should be warned about the phenomenon called "Illusory Truth Effect". The illusory truth effect describes how we automatically assign higher truth value to information that we encounter multiple times. This entails that when we encounter information more than once, we are more likely to think that this information is true. For this part of the experiment, you should try to avoid the illusory truth effect as good as you can. Please try to make your judgement unaffected by illusory truth.”

Then, participants were instructed to rate the following statements for how truthful they find them using a 7-point Likert scale, ranging from “absolutely untrue” to “absolutely true”. Part of the instruction was also that participants were supposed to make use of the full scale. For this task, all statements from the re-exposure phase and new statements were rated. Each statement was shown only for 15 s, to prevent participants from looking up answers on the internet. The test phase consisted of 100 trials, with five different sets of statements containing 20 statements each. The different sets of statements were 1) statements shown in both initial exposure phase and during the non-retrieval task, here named “Baseline Old” 2) statements shown only during the non-retrieval task, here named “Baseline New”, 3) statements shown in both initial exposure phase and during the retrieval practice, here named “Retrieval Old” 4) statements shown only during the retrieval practice, here named “Retrieval New” and 5) new statements shown only during the test phase, here named “Test Control”.

After the experiment finished, participants were redirected to Qualtrics for a debrief.

Measures

Truth Rating

The dependent variable that is being measured in this study is the truth rating that people are giving in the test phase of the experiment. Statements are rated on a 7-point Likert scale and the steps are “absolutely untrue” (1), “moderately untrue”, “somehow untrue”, “neutral”, “somehow true”, “moderately true” and “absolutely true” (7).

Repetition

Repetition is an independent variable and refers to the number of times a statement is presented in the experiment. The statements in the “Retrieval Old” and “Baseline Old” conditions are presented three times, statements in the “Retrieval New” and “Baseline New” conditions are presented two times, and the “Test Control” statements are shown once.

Truth Status

The objective truth of a statement is reflected in the Truth Status variable, which is independent. The truth status is either “True” or “False”.

Intervention

The intervention variable refers to the different intervention tasks used in the training phase. The different tasks are either the retrieval practice task, or the non-retrieval task.

Difficulty

The difficulty of a statement is based on the recall norming from Wertgen and Richter (2023). The norming values in their study describe what percentage of participants were able to recall the answer to the question at hand. That would mean that a high recall norming indicates that a question was easy, and a lot of participants knew the

answer. In this analysis, however, the recall norming is transformed to reflect difficulty using the following transformation more accurately:

$$Difficulty = 1 - Recall\ norming$$

This way, a higher difficulty means that less participants in the norming study by Wertgen and Richter (2023) were able to recall the answer to the question.

Reaction Time

Reaction Time used in this study refers to the time (s) it took participants to answer each item in the truth rating task during the final test phase. In a classic item recognition task, fluently processed stimuli are accompanied by shorter identification reaction times (Stark & McClelland, 2000). Unkelbach and Rom (2017) also demonstrated shorter response latencies for repeated statements than for new statements. Therefore, reaction time will be used as a marker for processing fluency (Huang & Shanks, 2021).

Interest Rating

Interest Rating refers to the results of the rating task in the study phase. Interest was rated using a 7-point Likert scale with the steps “not interesting”, “2”, “3”, “4”, “5”, “6”, and “very interesting”. It is important to note that there is no negative end of the scale. This was deliberately chosen, as piloting experiences have shown that it is hard for participants to assign a negative interest to something, meaning there is not really anything “very uninteresting”.

Recollection Success

The recollection success is an independent variable and reflects whether participants successfully identified a stimulus as old or new during the retrieval practice task in the re-exposure phase. If a statement is correctly identified, it is a “hit”, otherwise a “miss”.

Analysis

Mixed-effects modeling was used to address each hypothesis individually. This type of analysis was employed as mixed-effects modeling holds the advantage that

each participant can be included as a random effect, as well as being able to account for important sources of variance in the data (Baayen et al., 2008). The analysis of the obtained data was performed in R Software (R Core Team, 2023). Summary statistics were obtained using the *rstatix* package (Kassambara, 2023) and mixed-effects modelling was performed using the *lme4* package (Bates et al., 2015). After the initial summary statistics were obtained, the variables truth rating and interest rating were standardized before the analysis.

For Model 1, the presence of an illusory truth effect was investigated. Truth rating was predicted using repetition and truth status of each stimulus as fixed effects. For Model 2, the different Intervention tasks were additionally included to determine if there is an effect of retrieval practice on the illusory truth effect. Model 3 and Model 4 were built to further explore reaction time, which served as a marker for processing fluency. Furthermore, the effect of difficulty was accounted for. Finally, Model 5 was built to examine the relationship between recollection success during the retrieval practice on processing fluency in the truth rating. As the data from the non-retrieval task does not allow any analysis of memory performance, the data was reduced to exclude statements from the non-retrieval task. Then, recollection success in the retrieval practice was used to predict reaction time during the truth rating.

For each model, a null model was fitted. Furthermore, each participant was included as a random effects predictor which was restrained to an intercept. Nesting the random effect of participants within the random effect education area did not provide any significant explanatory power to any model, as indicated by a likelihood ratio test (χ^2 = 9.3, *df* = 14, *p* = .812), and was therefore not included in the final models to avoid overfitting. Furthermore, items were included as a random effect.

Ethical Considerations

We did not expect any negative consequences to the participants that could have been caused by the experiment. Participants were briefed in detail about their right to withdraw from the experiment at any point. Furthermore, participants were asked to provide consent to participate in the study, and only consenting participants were able to start the experiment. The data collected as part of this study was fully

anonymized and answers could not be traced back to any participant, which was also communicated to the participants. Overall, the study was carried out in accordance with guidelines provided in the Swedish Ethical Review Act (Swedish Ethical Review Authority, 2023).

Open Science

In an effort to contribute to transparency and reproducibility in research, all materials used in this study are accessible on the Open Science Framework (https://osf.io/6uecj/?view_only=476697b0356e4fa3b293c6b4857bd918).

Results

Sample

Table 1 shows descriptive statistics for the collected data, where the mean truth rating was slightly above "neutral" at *M* = 4.41 (*SD* = 2.1). Interest ratings were slightly below the middle at *M* = 3.45 (*SD* = 1.74). The mean hit rate was 89% (*M* = .89) for the retrieval practice (*SD* = .13) and 74% (*M* = .74) for the categorization task (*SD* = .15). The mean reaction time for the truth rating was *M* = 3.9 s (*SD* = 1.22).

Table 1

Summary Statistics

Variable	*N*	*M*	*SD*
Age	87	27.1	8.92
Truth Rating	8502	4.41	2.1
Interest Rating	3268	3.45	1.74
Hit Rate Retrieval Practice	88	.89	.13
Hit Rate Categorization Task	88	.74	.15
Mean Reaction Time (s)	88	3.9	1.22

Note. This table shows descriptive statistics for the obtained data. *N* indicates the numbers of observation per variable, *M* shows the mean score, and *SD* shows the standard deviation.

Illusory Truth Effect

To test whether an Illusory Truth Effect occurred, Model 1 was fitted using truth rating as the outcome variable. Fixed effect predictors included the repetition and the truth status. The model also included an interaction between repetition and truth status.

Table 2

Table of Coefficients for Model 1

	B	95% CI		β	*p*-value
		LL	*UL*		
Intercept (Control)	-.7	-.83	-.45	0	< .001 *
Repetition	.09	.06	.12	.07	< .001 *
Truth Status	1.07	.9	1.25	.54	< .001 *
Interaction Rep x Truth Status True	-.04	-.09	0	-.05	.065

Note. This table shows significant predictors for Model 1. The regression coefficient is shown by *B*, 95% confidence intervals are shown including the lower limit (*LL*) and the upper limit (*UL*). Furthermore, standardized beta-coefficient is indicated by *β*. The intercept represents statements that had a repetition value of 1, meaning that they were only shown once. Significance level of *p*-values are *p < .05.

As seen in Table 2, there was a very small illusory truth effect (β = .07 [95% CI = .06, .12], $p < .001$), where repeated statements were rated more true than new statements. Each repetition was accompanied by an increase in truth ratings by .09 standard deviations. Furthermore, there was a large effect of truth status ($B = 1.07$ [95% CI = .9, 1.25], $p < .001$), as true statements where rated higher than false statements (β = .54). All fixed effect predictors combined were able to explain 24% of the variance in the outcome variable ($R^2 = .24$). The interaction between repetition and truth status almost reached significance ($B = -.04$ [95% CI = -.09, 0], $p = .065$), where the illusory truth effect was weaker for true statements than for false statements (β = -.04). However, that relationship should be evaluated critically, as confidence intervals include zero (95% CI = -.09, .00).

Retrieval Practice

Model 2 was built to examine whether there is an effect for an interaction between intervention and repetition status on the truth rating as outcome variable. In addition to Model 1, the intervention type was included as a fixed effect predictor. Furthermore, an interaction between repetition and intervention was added.

Table 3

Table of Coefficients for Model 2

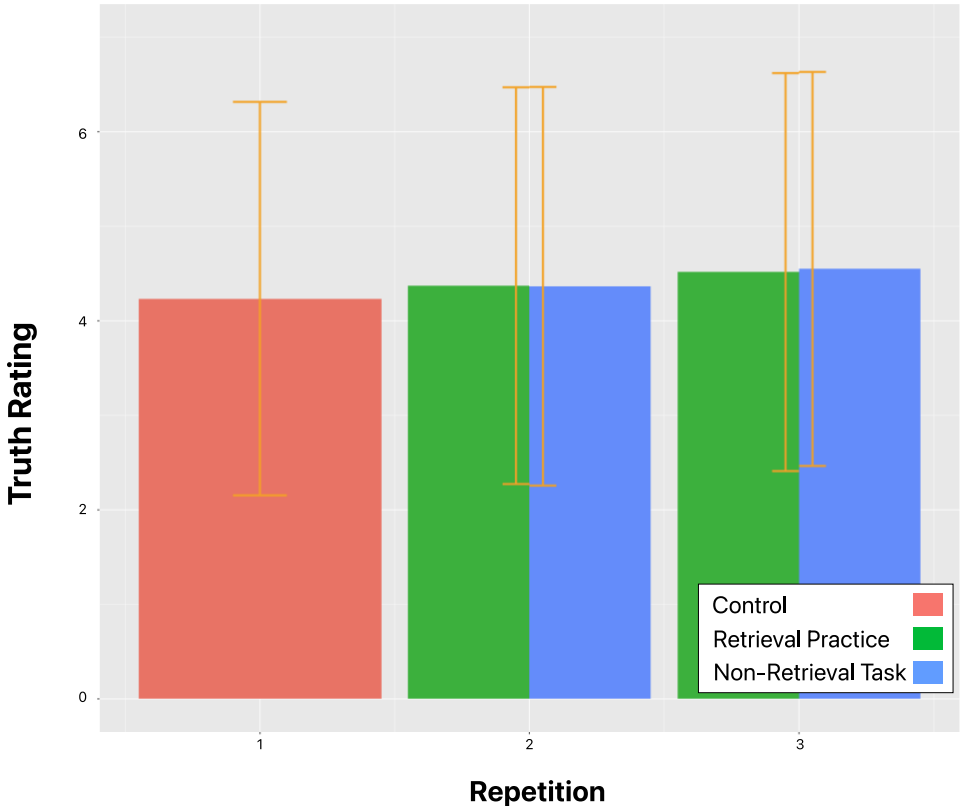
	B	95% CI		β	*p*-value
		LL	*UL*		
Intercept (Control)	-.69	-.83	-.55	0	< .001 *
Repetition	.09	.03	.15	.07	.002 *
Non-Retrieval Task	-.04	-.19	.11	-.02	.595
Retrieval Practice	0	-.09	.09	0	.962
Truth Status True	1.07	.9	1.25	.54	< .001 *
Interaction Rep x Non-Retrieval Task	.01	-.06	.09	.02	.721
Interaction Rep x Truth Status True	-.04	-.09	0	-.05	.065

Note. This table shows significant predictors for Model 2. The regression coefficient is shown by *B*, 95% confidence intervals are shown including the lower limit (*LL*) and the upper limit (*UL*). Furthermore, standardized beta-coefficient is indicated by *β*. The intercept represents statements that had a repetition value of 1, meaning that they were only shown once. Significance level of *p*-values are *p < .05.

As seen in Table 3, there was a very small effect of repetition (*B* = .09 [95% CI = .03, .15], *p* = .002) and a large effect of truth status (*B* = 1.07 [95% CI = .9, 1.25], *p* < .001). Again, each repetition was accompanied by an increase in truth ratings by .09 standard deviations. All fixed effect predictors combined were able explain 24% (R^2 = .24) of the variance of the outcome variable (see Appendix A for full table of marginal R^2). Similar to Model 1, true statements were rated higher than false statements (β = .54) and truth ratings increased with repetition (β = .07). Figure 2 shows mean truth ratings across different repetitions and intervention types. The illusory truth effect was not affected by retrieval practice, as there was no significant interaction between repetition and the intervention with confidence intervals including zero (*B* = .01 [95% CI = -.06, .09], *p* = .721).

Figure 2

Mean Truth Rating by Repetition Status and Intervention Type



Note. This bar plot shows the mean truth rating by repetition status. Furthermore, green bars represent statements that were part of the memory task, and blue bars represent statements from the categorization task. Generally, mean truth ratings increase with increasing repetition. The difference between the memory task and the category task is minimal.

Furthermore, after accounting for repetition status, the interaction between intervention and repetition status did not add any significant amount of explanatory power to the model as marginal r squared was R^2 = .001 for repetition and R^2 < .0001 for the interaction between decoding intervention and repetition. The main effects for the individual intervention tasks seem negligible as well, as confidence intervals are wide and include zero for both retrieval practice (*B* = 0 [95% CI = -.09, .09], *p* = .962) and the non-retrieval task (*B* = -.04 [95% CI = -.19, .11], *p* = .592).

Processing Fluency

For Model 3, a model was fitted using the truth rating as outcome variable. Fixed effect predictors were truth status, repetition, reaction time and difficulty. Furthermore, interaction effects between reaction time and truth status, between

reaction time and repetition as well as between truth status and difficulty were included.

Table 4
Table of Coefficients for Model 3

	B	95% CI		β	*p*-value
		LL	*UL*		
Intercept (False)	-.67	-.92	-.42	0	< .001 *
Truth Status True	1.48	1.13	1.82	.74	< .001 *
Repetition	.1	.05	.14	.07	< .001 *
Reaction Time	-.02	-.04	0	-.05	.073
Difficulty	.27	-.13	.67	.06	.197
Interaction Truth Status True x Reaction Time	.02	0	.03	.05	.012 *
Interaction Repetition x Reaction Time	-.01	-.02	0	-.06	.041 *
Interaction Truth Status True x Difficulty	-1.2	-1.84	-.57	-.33	< .001 *

Note. This table shows significant predictors for Model 3. The regression coefficient is shown by *B*, 95% confidence intervals are shown including the lower limit (*LL*) and the upper limit (*UL*). Furthermore, standardized beta-coefficient is indicated by *β*. The intercept represents statements that had a truth status of “False”, meaning they were objectively false. Significance level of *p*-values are *p < .05.

As seen in Table 4, there was a large effect for truth status (*B* = 1.48 [95% CI = 1.13, 1.82], *p* < .001). True statements were rated more true than false statements (β = .74). There was also a very small illusory truth effect (*B* = .1 [95% CI = .05, .14], *p* < .001), where repeated statements were rated more true than new statements (β = .07). Each repetition was accompanied by an increase in truth ratings by .1 standard deviations. All fixed effect predictors combined were able explain 27% (R^2 = .27) of the variance of the outcome variable. The interaction between truth status and reaction time (*B* = .02 [95% CI = 0, .03], *p* < .012) should be interpreted carefully, as the confidence intervals are including zero. Generally, the effect of reaction time on truth ratings was stronger for true statements (β = .05) than for false statements. Moreover, the interaction between repetition and reaction times reached significance (*B* = -.01 [95% CI = -.02, 0], *p* = .041), where the illusory truth effect was smaller the longer people took to answer (β = -.06). However, this effect should be interpreted

careful as confidence intervals include zero (95% CI = -.02, 0). Finally, there was a medium effect for the interaction between truth status and difficulty (*B* = -1.2 [95% CI = -1.84, -.57], *p* < .001), where the effect of truth status was weaker for hard statements (β = -.33) than for easy statements. The effect for statement difficulty was not significant (*B* = .27 [95% CI = -.13, .67], *p* = .197).

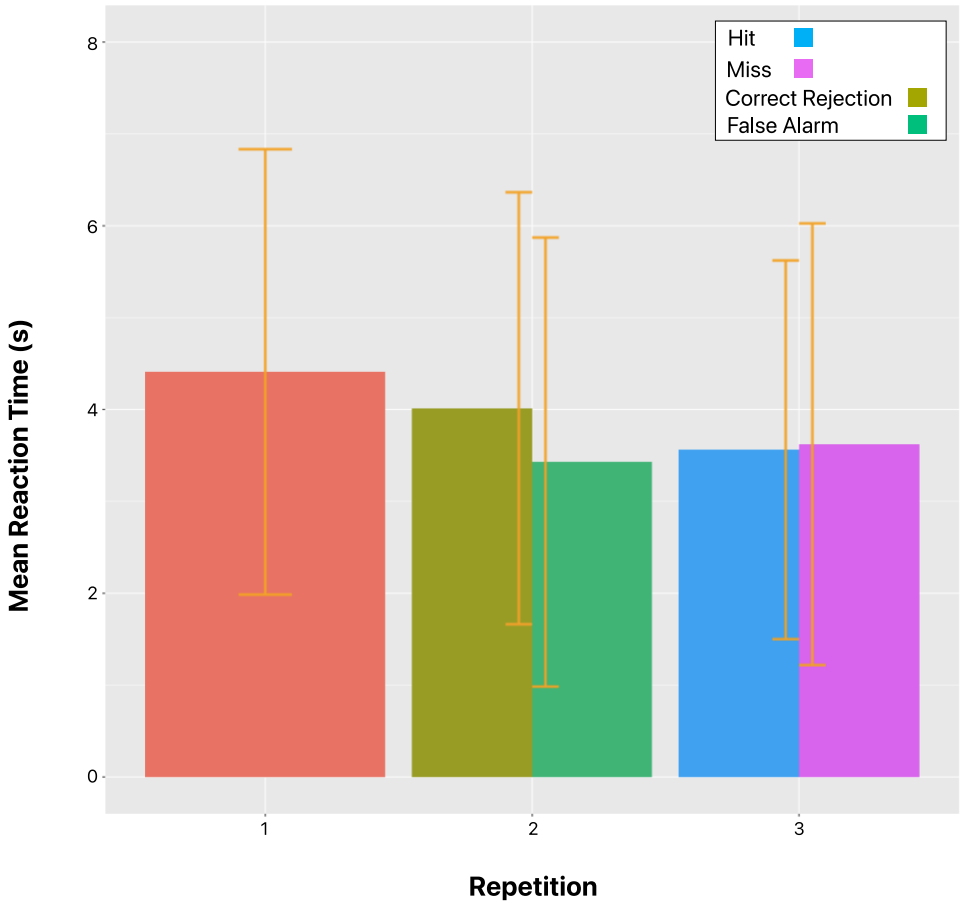
For model 4, reaction time was predicted using repetition, intervention, and difficulty as fixed effect predictors. Furthermore, interaction terms between repetition and difficulty, as well as between repetition and intervention were included. There was a medium effect for repetition (*B* = -.63 [95% CI = -.85, -.41], p < .001) and a small effect for difficulty (*B* = -1.38 [95% CI = -2.11, -.64], p < .001) (see Appendix B for full table of coefficients). Generally, reaction times were shorter for repeated statements than for new statements (β = -.21) as well as for harder statements (β = -.13). Each repetition was accompanied by a decrease in reaction time of .63 s. All fixed effect predictors combined were able explain 2.3% (R^2 = .02) of the variance of the outcome variable. Furthermore, there was a significant interaction between repetition and difficulty (*B* = .46 [95% CI = .05, .87], p = .029), as with increasing difficulty, the effect of repetition on the reaction times increased (β = .12). The effect of the retrieval practice did not reach significance (*B* = .25 [95% CI = -.17, .67], p = .249), nor did the effect for the non-retrieval task (*B* = .18 [95% CI = -.25, .6], *p* = .412).

For Model 5, recollection success and repetition were used to predict reaction time in the final truth rating phase. There was a small effect for repetition (*B* = -.45 [95% CI = .58, -.32], *p* < .001) where repeated statements were answered faster than non-repeated ones (β = -.1) (see Appendix B for full table of coefficients). Each repetition was accompanied by a decrease in reaction time by .45 s. All fixed effect predictors combined were able explain .9% (R^2 = .009) of the variance of the outcome variable. Furthermore, there was a significant interaction between recollection success and repetition (*B* = .5 [95% CI = .07, .93], *p* = .022), as the effect of repetition was stronger, when a statement was not correctly identified during retrieval practice (β = .19). The main effect of recollection success almost reached significance (*B* = -1.15 [95% CI = -2.3, 0]), *p* = .05), where the reaction time was shorter for statements that were not correctly identified during the retrieval practice (β = -.16). Figure 3 shows mean reaction times during the final truth rating ordered by repetition and the

recollection success in the retrieval practice. Depending on the repetition status, the relationship between recollection success and reaction time changed. For old statements in the retrieval practice, hits had shorter mean reaction time (*M* = 3.56) than misses (*M* = 3.62). For new statements, the relationship changed, where hits had a longer mean reaction time (*M* = 4.01) than misses (*M* = 3.43).

Figure 3

Mean Reaction Time by Repetition and Recollection Success



Note. This figure shows mean reaction time during the final truth rating task ordered by repetition and recollection success in the retrieval practice. For new statements in the retrieval practice, hits are referred to as “correct rejections” and misses are referred to as “false alarms”. For old statements, hits had shorter mean response times than misses. However, for new statements, correct rejections had longer reaction times than false alarms.

Discussion

This study investigated whether retrieval practice can suppress the illusory truth effect. Participants rated statements for subjective truth. There was a very small illusory truth effect on the final truth ratings, with repeated statements being rated more true than new statements. The illusory truth effect that appeared in this study is

much smaller than the average effect reported in the meta-analysis by Dechêne et al. (2010), who reported an average medium effect size (d = .32 to d = .55). Repeated statements were also answered faster than new statements, which suggests that those items were processed more fluent than new items. The actual, objective truth value of the statements had the largest impact on the final truth ratings, with true statements being rated much higher than false statements.

Retrieval practice was not able to reduce the illusory truth effect differentially, as there were no effects for the interaction between retrieval practice and repetition. This relationship also did not change when taking the truth status into account, i.e., it did not differ between true and false statements. There was also no interaction between the non-retrieval task and repetition. However, the illusory truth effect seen in this study was very small, suggesting that while retrieval practice does not enhance the suppression of illusory truth differentially, there might be a joint effect for both intervention tasks. Furthermore, the intervention types did not have any effect on reaction times in the final truth rating phase, which could have been a sign for differences in processing fluency between the memory and the categorization task.

Explanations for the Size of the Illusory Truth Effect

The small illusory truth effect might be due to warning before the test phase and attest that it works as intended. Other studies that used a similar warning yielded illusory truth effects of medium to small effect sizes. The illusory truth effect in the study by Jalbert et al. (2020) had an effect size of d = .72 for statements that included a warning and Calio et al. (2020) reported an effect size of d_z = .27. As the results of this study indicate a much smaller illusory truth effect (β = .07), possible explanations of this will be discussed in the following.

Levels of processing during the exposure phase might have influenced the size of the illusory truth effect in this study. The illusory truth effect should be the highest for the statements with the most repetitions, i.e., the “Retrieval Old” and “Baseline Old” conditions. Their initial exposure task was an interest rating, similar to the task that Brashier et al. (2020) used. Dechêne et al. (2010) present in their moderator analysis that the illusory truth effect decreases with higher levels of processing during the first

exposure. Using the framework proposed by Craik and Lockhart (1972), we can establish that self-referential processing belongs to the deeper levels of processing, so the interest rating task should fall under that category as well. However, the finding that level of processing decreases the illusory truth effect have been opposed by Unkelbach and Rom (2017), who found that self-referential processing actually increases the illusory truth effect. Brashier et al. (2020) also found a large illusory truth effect for statements initially encoded with an interest rating. In the light of those findings, it is reasonable to assume that the initial interest task in this study did not decrease the illusory truth effect to begin with.

The second phase of the experiment included a retrieval practice task and a non-retrieval task, where the retrieval practice did not seem to affect the illusory truth effect differentially. Considering the case of the non-retrieval task, the findings can be compared to those by Nadarevic and Erdfelder (2014), where statements initially encoded by assigning them to a knowledge category led to a large illusory truth effect. We must draw the important distinction that the task in this study was a secondary, re-exposure phase, and therefore the results can only be compared to a small extent. In this study, the attempt was to affect the illusory truth effect after the first exposure already happened. The results also open the possibility that the initial encoding of information could be the dominating factor in how the information is represented and secondary exposure stages might not be strong enough to overturn this representation.

Other factors that might have reduced the illusory truth effect from the start is that truth ratings were obtained using a 7-point Likert-Scale, which yielded the smallest truth effect out of the rating scales compared to even scales or dichotomous response formats (Dechêne et al., 2010).

Furthermore, the study was conducted on a computer, and that the repeated exposure and final truth rating happened on the same day. Both of those factors have been identified as yielding smaller illusory truth effects, than pen and paper tests and a longer interval between prior exposure and final truth rating, respectively (Dechêne et al., 2010). To sum up, what we might have seen here was the influence of many moderating effects that decrease the illusory truth effect.

Successful Recollection

For the statements that were part of the retrieval practice, recollection success had a moderating influence on whether repetition of statements enhanced processing fluency. For old statements that were identified correctly during the retrieval practice, reaction times during the truth rating were shorter than for statements that were missed. For new statements that were identified as such, reaction times were longer than for statements that were missed. Considering the findings of Guran et al. (2020), we would not expect to see that the effectiveness of the retrieval practice depends on whether a statement is remembered correctly in the retrieval phase. In their study, there was no reported interaction between recollection success during the retrieval practice and recollection performance during the final memory task (Guran et al., 2020). It is also important to mention that the comparison of “hit”-, and “miss”-items is based on very different sample sizes, as the average rate of hits during the retrieval practice was 89%.

However, the results resemble the findings of Huang and Shanks (2021), who found that successful recollection of source memory is associated with higher processing fluency. In their study, hits were associated with the lowest reaction time, and correct rejections had the highest reaction times (Huang & Shanks, 2021). There was a similar pattern seen in the present study. However, there was no direct influence of processing fluency on truth ratings. This suggests that recollection success might have been accompanied by higher fluency, however this did not affect the illusory truth effect in any way. This opens the interpretation that occurrence of fluency alone is not the determining factor driving the illusory truth effect, but the misattribution to truth.

Considering the effect of processing fluency hints towards a more complex relationship between recollection, processing fluency and the illusory truth effect. The interaction between reaction time and repetition did not reach significance. This means that processing fluency does not reliably predict the strength of the illusory truth effect. Furthermore, the intervention type did not influence processing fluency, indicated by reaction times in the final truth rating phase. The results can be

interpreted towards the idea that recollection alone does not reduce the illusory truth effect. It is possible that even though recollection was enhanced, processing fluency was still not successfully attributed to the recollected exposure, and the fluency experience was not discounted. Future interventions should therefore focus on the process of attribution and assist people in experiencing and discounting the feeling of processing fluency.

Difficulty

Statement difficulty has impacted the reaction times of participants in this study. If we use this as evidence for fluent processing, the results suggest that difficult statements were processed more fluent than easy statements. Furthermore, this effect was stronger for repeated items. Although it seems counterintuitive that difficult statements are processed more fluent, it allows for the interpretation that when participants had no relevant knowledge to use as a cue in their judgement process, that they relied strongly on processing fluency instead.

Furthermore, difficulty affected the impact that the objective truth value of a statement had on the final truth ratings. This goes to show that the norming of the statements, conducted by Wertgen and Richter (2023), provided a good estimate for which statements participants likely held relevant knowledge, and which not. As statements became easier, the objective truth value was able to explain most of the variance in the outcome variable (see Appendix C). The opposite applied to harder questions, where objective truth value lost its explanatory power, the harder the questions became. This indicates that participants used their knowledge when present, however relied less and less on knowledge when statements became harder.

Other Factors Influencing Truth Decisions

A common theme throughout the analysis of this study was that, after including all of our fixed and random effects, there was still a lot of variance (> 70%) unaccounted for. This is a testament to the many factors that influence decisions about truth. As Brashier and Marsh (2020) summarize, truth judgements are informed by multiple cues, where factual knowledge is only one of many. For example, there is

an inherent bias towards judging statements as true rather than false, as in the real world, we usually encounter information that is true, and people tend to tell the truth. Furthermore, the source of a statement has a huge impact on whether we trust it to be true or not. Even cues like feelings of affect can influence what we believe to be true (Brashier & Marsh, 2020).

There is also a personal and political factor involved in the evaluation of information, where affirmation of own worldviews increases the belief in an information (Ecker et al., 2022). The statements used in this study are chosen to be politically neutral, however, real-world applications of an intervention as shown in this study would need to consider political content as well.

Limitations

The results of this study need to be evaluated considering several limitations that come with it. The sample might have been a biasing factor, with most of the sample sourced from the platforms SurveySwap and Reddit. Due to the nature of SurveySwap, participants are mainly other students or researchers who are looking for participants for their own studies, so the sample will consist of those. However, this should not be a reason of concern, as the vast majority of illusory truth studies were actually carried out using students as a sample (Henderson et al., 2022). Furthermore, the study was advertised on the Reddit community “r/SampleSize”, which is a place for researchers to advertise their studies and gain participants. Likely, participants who choose to do the study are interested in the research topic already and might be familiar with the concept of the illusory truth effect. To avoid this, the study was only advertised as a study on “information processing” and no further information was given. However, given that people have started but not completed the experiment, it cannot be ruled out that interest in the research topic has affected who completed the full study.

Considering that, to our knowledge, the use of retrieval practice in a study investigating the illusory truth effect has not been attempted before, it is not guaranteed that the same intervention effects seen in Guran et al. (2020) transfer to the setup and material used in this study. Moreover, Brashier et al. (2020) argue that

their intervention only worked when participants had relevant knowledge about the statements to rate, and that their intervention encouraged participants to use that knowledge. We cannot rule out that in the present study, even when participants were able to recollect an item from a previous phase, they had no relevant knowledge for the item at hand, and therefore resorted to using processing fluency as a cue for their truth judgement.

Another limitation is that there is no clear way to tell if the retrieval practice intervention did in fact increased recollection performance in the truth rating phase. There were no memory checks administered to control if participants remember a statements from a previous round. As this was done mainly due to limit the time it takes to do the experiment, retrospectively, this could have provided valuable insights into why the retrieval practice did not differentially affect the illusory truth effect.

Finally, there was no control condition for participants being presented with no warning at all. This does not allow for an interpretation if the illusory truth effect was smaller compared to having no warning. The focus on this study was to investigate if the illusory truth effect could be suppressed by retrieval practice, and the control task used in this study were statements being categorized. However, we cannot exclude the possibility that the categorization task had its unique effect that matched the retrieval practice in its effectiveness to reduce the illusory truth effect.

Implications

The results of this study can be used to further inform the way we handle the fight against misinformation and fake news. Educating the public about the role of the illusory truth effect and its impact on truth judgement is a good first step. However, to eliminate it, stronger interventions that target fluency misattributions at the cognitive level might be the answer. Furthermore, this study can help drive the discussion about factors influencing the illusory truth effect, and further acknowledge the need for interventions helping people to discount the feeling of fluency.

Another issue that needs to be addressed is that we usually do not have control about the first exposure to information, meaning that in a real-world setting, the first

exposure has already happened. This is a roadblock for traditional fact-checking and debunking efforts, as repeating the initial false information to correct it increases repetition and familiarity, causing the debunking effort to backfire (Ecker et al., 2020; Swire et al., 2017). To effectively target misinformation, interventions need to take place at the retrieval stage, allowing the reduction of the illusory truth effect after the information is already encoded.

Future Research

This study does leave some open questions that future research might latch on to. In this study, participants were better at distinguishing true and false statements, when they remembered a statement during retrieval practice. In future studies, this relationship could be evaluated in depth, perhaps by introducing memory checks during the final truth rating phase.

Furthermore, as a lot of information is consumed through social media nowadays (Aïmeur et al., 2023), it would be interesting testing the effects of warnings and retrieval practice in the social media ecology. Furthermore, the results of this study should be tested for its robustness by replicating this study with a larger, more representative sample, including people from different age groups and educational backgrounds. Furthermore, it would be interesting to see if the results replicate using different stimuli contents, like news headlines, as in Calvillo and Smelter (2020) ,but also new and more ecologically valid content formats, e.g. images and videos.

Finally, the design of the experiment does not allow any prediction about what long term effects the intervention might bring. In a realistic setting, the different exposures to an information that cause the illusory truth effect do not immediately follow each other. Therefore, it would be worthwhile investigating the role of recollection on suppressing the illusory truth effect in a longitudinal study.

Conclusion

In this study, retrieval practice was used to suppress the illusory truth effect. Participants were exposed to trivia statements in three different phases including a

final truth rating. The results suggest that the retrieval practice used in this study did not differentially suppress the illusory truth effect. It is possible that participants were able to recollect the prior exposure but chose to resort to processing fluency, as the exposure was not attributed to fluency and therefore discounted. Another explanation could be that participants resorted to fluency, as there were no other relevant cues to base the truth decision on. Furthermore, the non-retrieval task used as a control might have had its own unique influence on the outcome variable. The results also open the possibility that the initial exposure to an information has a large impact on the later illusory truth effect, and that interventions afterwards only have a small impact on it.

The findings of this study can advance the discussion about the illusory truth effect and how to avoid it. Helping people be aware of the fluency experience, and correctly attributing fluency to prior exposure should be the focus of future interventions. The final goal should be to find effective means against the spread and belief in misinformation, as the negative consequences are severe and affect many domains of our life.

References

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, *13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Sci Adv*, *6*(14), eaay3539. https://doi.org/10.1126/sciadv.aay3539

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219-235. https://doi.org/10.1177/1088868309341564

Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, *2*(2), 81-94. https://doi.org/https://doi.org/10.1002/bdm.3960020203

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1 - 48. https://doi.org/10.18637/jss.v067.i01

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*, 446-458. https://doi.org/10.1037/0096-3445.121.4.446

Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, *194*, 104054. https://doi.org/10.1016/j.cognition.2019.104054

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*(1), 499-515. https://doi.org/10.1146/annurev-psych-010419-050807

Calio, F., Nadarevic, L., & Musch, J. (2020). How explicit warnings reduce the truth effect: A multinomial modeling approach. *Acta Psychologica*, *211*, 103185. https://doi.org/10.1016/j.actpsy.2020.103185

Calvillo, D. P., & Smelter, T. J. (2020). An initial accuracy focus reduces the effect of prior exposure on perceived accuracy of news headlines. *Cognitive Research: Principles and Implications*, *5*(1), 55. https://doi.org/10.1186/s41235-020-00257-y

Corneille, O., Mierop, A., & Unkelbach, C. (2020). Repetition increases both the perceived truth and fakeness of information: An ecological account. *Cognition*, *205*, 104470. https://doi.org/10.1016/j.cognition.2020.104470

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671-684. https://doi.org/10.1016/S0022-5371(72)80001-X

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238-257. https://doi.org/10.1177/1088868309352251

Dew, I. T., & Cabeza, R. (2013). A broader view of perirhinal function: from recognition memory to fluency-based decisions. *J Neurosci*, *33*(36), 14466-14474. https://doi.org/10.1523/jneurosci.1413-13.2013

Ecker, U. K. H., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), 41. https://doi.org/10.1186/s41235-020-00241-6

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29. https://doi.org/10.1038/s44159-021-00006-y

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annu Rev Neurosci*, *30*, 123-152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Garcia-Marques, T., Silva, R. R., & Mello, J. (2016). Judging the truth-value of a statement In and out of a deep processing context. *Social Cognition*, *34*(1), 40-54. https://doi.org/10.1521/soco.2016.34.1.40

Guran, C.-N. A., Lehmann-Grube, J., & Bunzeck, N. (2020). Retrieval Practice Improves Recollection-Based Memory Over a Seven-Day Period in Younger and Older Adults [Original Research]. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02997

Hansen, J., Dechêne, A., & Wänke, M. (2008). Discrepant fluency increases subjective truth. *Journal of Experimental Social Psychology*, *44*(3), 687-691. https://doi.org/10.1016/j.jesp.2007.04.005

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107-112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hawkins, S. A., & Hoch, S. J. (1992). Low-Involvement Learning: Memory without Evaluation. *Journal of Consumer Research*, *19*(2), 212-225. https://doi.org/10.1086/209297

Henderson, E. L., Westwood, S. J., & Simons, D. J. (2022). A reproducible systematic map of research on the illusory truth effect. *Psychonomic Bulletin & Review*, *29*(3), 1065-1088. https://doi.org/10.3758/s13423-021-01995-w

Huang, T. S.-T., & Shanks, D. R. (2021). Examining the relationship between processing fluency and memory for source information. *Royal Society Open Science*, *8*(4), 190430. https://doi.org/10.1098/rsos.190430

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*, 306-340. https://doi.org/10.1037/0096-3445.110.3.306

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*(3), 326-338. https://doi.org/10.1037/0022-3514.56.3.326

Jalbert, M., Newman, E., & Schwarz, N. (2020). Only half of what I'll tell you is true: Expecting to encounter falsehoods reduces illusory truth. *Journal of Applied Research in Memory and Cognition*, *9*(4), 602-613. https://doi.org/10.1016/j.jarmac.2020.08.010

Kassambara, A. (2023). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. In (Version R package version 0.7.2) https://rpkgs.datanovia.com/rstatix/

Kruijt, J., Meppelink, C. S., & Vandeberg, L. (2022). Stop and Think! Exploring the Role of News Truth Discernment, Information Literacy, and Impulsivity in the Effect of Critical Thinking Recommendations on Trust in Fake Covid-19 News. *European Journal of Health Communication*, *3*(2), 40-63. https://doi.org/10.47368/ejhc.2022.203

Lazer, D. M. J., 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. https://doi.org/10.1126/science.aao2998

Lewandowsky, S., Ecker, U. K. H., & 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. https://doi.org/10.1016/j.jarmac.2017.07.008

Li, B., Gao, C., Wang, W., & Guo, C. (2015). Processing fluency hinders subsequent recollection: an electrophysiological study [Original Research]. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00863

Mitchell, J. P., Dodson, C. S., & Schacter, D. L. (2005). FMRI evidence for the role of recollection in suppressing misattribution errors: The illusory truth effect. *Journal of Cognitive Neuroscience*, *17*(5), 800-810. https://doi.org/10.1162/0898929053747595

Nadarevic, L., & Aßfalg, A. (2017). Unveiling the truth: warnings reduce the repetition-based truth effect. *Psychological Research*, *81*(4), 814-826. https://doi.org/10.1007/s00426-016-0777-y

Nadarevic, L., & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, *23*, 74-84. https://doi.org/10.1016/j.concog.2013.12.002

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*, 1865-1880. https://doi.org/10.1037/xge0000465

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388-402. https://doi.org/10.1016/j.tics.2021.02.007

Pérez-Escolar, M., Lilleker, D., & Tapia-Frade, A. (2023). A Systematic Literature Review of the Phenomenon of Disinformation and Misinformation [credibility; disinformation; fake news; falsehood; hoaxes; misinformation; truth]. *2023*, *11*(2), 12. https://doi.org/10.17645/mac.v11i2.6453

Qualtrics. (2020). (Version July, 2020) Qualtrics. https://www.qualtrics.com

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338-342. https://doi.org/10.1006/ccog.1999.0386

Sadeh, T., Maril, A., & Goshen-Gottstein, Y. (2012). Encoding-related brain activity dissociates between the recollective processes underlying successful recall and recognition: A subsequent-memory study. *Neuropsychologia*, *50*(9), 2317-2324. https://doi.org/10.1016/j.neuropsychologia.2012.05.035

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns. In *Advances in Experimental Social Psychology* (Vol. 39, pp. 127-161). Academic Press. https://doi.org/10.1016/S0065-2601(06)39003-X

Skinner, E. I., & Fernandes, M. A. (2007). Neural correlates of recollection and familiarity: A review of neuroimaging and patient data. *Neuropsychologia*, *45*(10), 2163-2179. https://doi.org/10.1016/j.neuropsychologia.2007.03.007

Stark, C. E. L., & McClelland, J. L. (2000). Repetition priming of words, pseudowords, and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 945-972. https://doi.org/10.1037/0278-7393.26.4.945

Swedish Ethical Review Authority. (2023). *What the Act says*. https://etikprovningsmyndigheten.se/en/what-the-act-says/

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *J Exp Psychol Learn Mem Cogn*, *43*(12), 1948-1961. https://doi.org/10.1037/xlm0000422

Treen, K. M. d. I., Williams, H. T., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, *11*(5), e665. https://doi.org/10.1002/wcc.665

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 219-230. https://doi.org/10.1037/0278-7393.33.1.219

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In *The experience of thinking: How the fluency of mental processes influences cognition and behaviour.* (pp. 11-32). Psychology Press.

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, *28*(3), 247-253. https://doi.org/10.1177/0963721419827854

Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition*, *160*, 110-126. https://doi.org/10.1016/j.cognition.2016.12.016

van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460-467. https://doi.org/10.1038/s41591-022-01713-6

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Wang, W., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R. (2016). On known unknowns: Fluency and the neural mechanisms of illusory truth. *Journal of Cognitive Neuroscience*, *28*(5), 739-746. https://doi.org/10.1162/jocn_a_00923

Wertgen, A. G., & Richter, T. (2023). General knowledge norms: Updated and expanded for German. *PLOS ONE*, *18*(2), e0281305. https://doi.org/10.1371/journal.pone.0281305

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441-517. https://doi.org/10.1006/jmla.2002.2864

Zarocostas, J. (2020). How to fight an infodemic. *The lancet*, *395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Appendix A

Table A1
Marginal R^2 for Each Predictor of Model 2

	R^2	95% CI	
		LL	UL
Model	2.414e-01	2.23e-01	2.567e-01
Truth Status True	3.733e-02	2.989e-02	4.555e-02
Repetition	8.946e-04	7.568e-05	2.614e-03
Interaction Repetition x Truth Status True	3.110e-04	1.613e-06	1.511e-03
Non-Retrieval Task	2.184e-05	1.437e-07	7.120e-04
Interaction Repetition x Non-Retrieval Task	1.159e-05	1.275e-07	6.471e-04
Retrieval Practice	2.068e-07	1.157e-07	5.919e-04

Note. This table contains marginal R^2 for each predictor of Model 2. Truth status explains most of the variance in the outcome variable truth rating. Repetition explains only a small amount of variance and the additional variance explained by the interaction between repetition and truth status is minimal.

Appendix B

Table B1

Table of Coefficients for Model 4

	B	95% CI		β	*p*-value
		LL	*UL*		
Intercept (Control)	5.48	5.02	5.94	0	< .001 *
Repetition	-.63	-.85	-.41	-.21	< .001 *
Non-Retrieval Task	.18	-.25	.6	.04	.412
Retrieval Practice	.25	-.17	.67	.05	.249
Difficulty	-1.38	-2.11	-.64	-.13	< .001 *
Interaction Repetition x Difficulty	.46	0.05	.87	.12	.029 *
Interaction Retrieval Practice x Difficulty	-.32	-1.13	.48	-.04	.431
Interaction Non-Retrieval Task x Difficulty	-.57	-1.38	.23	-.07	.161

Note. This table shows predictors for Model 4. The regression coefficient is shown by *B*, 95% confidence intervals are shown including the lower limit (*LL*) and the upper limit (*UL*). Furthermore, standardized beta-coefficient is indicated by *β*. The intercept represents statements that had a repetition value of 1, meaning that they were only shown once. Significance level of *p*-values are *p < .05

Table B2

Table of Coefficients for Model 5

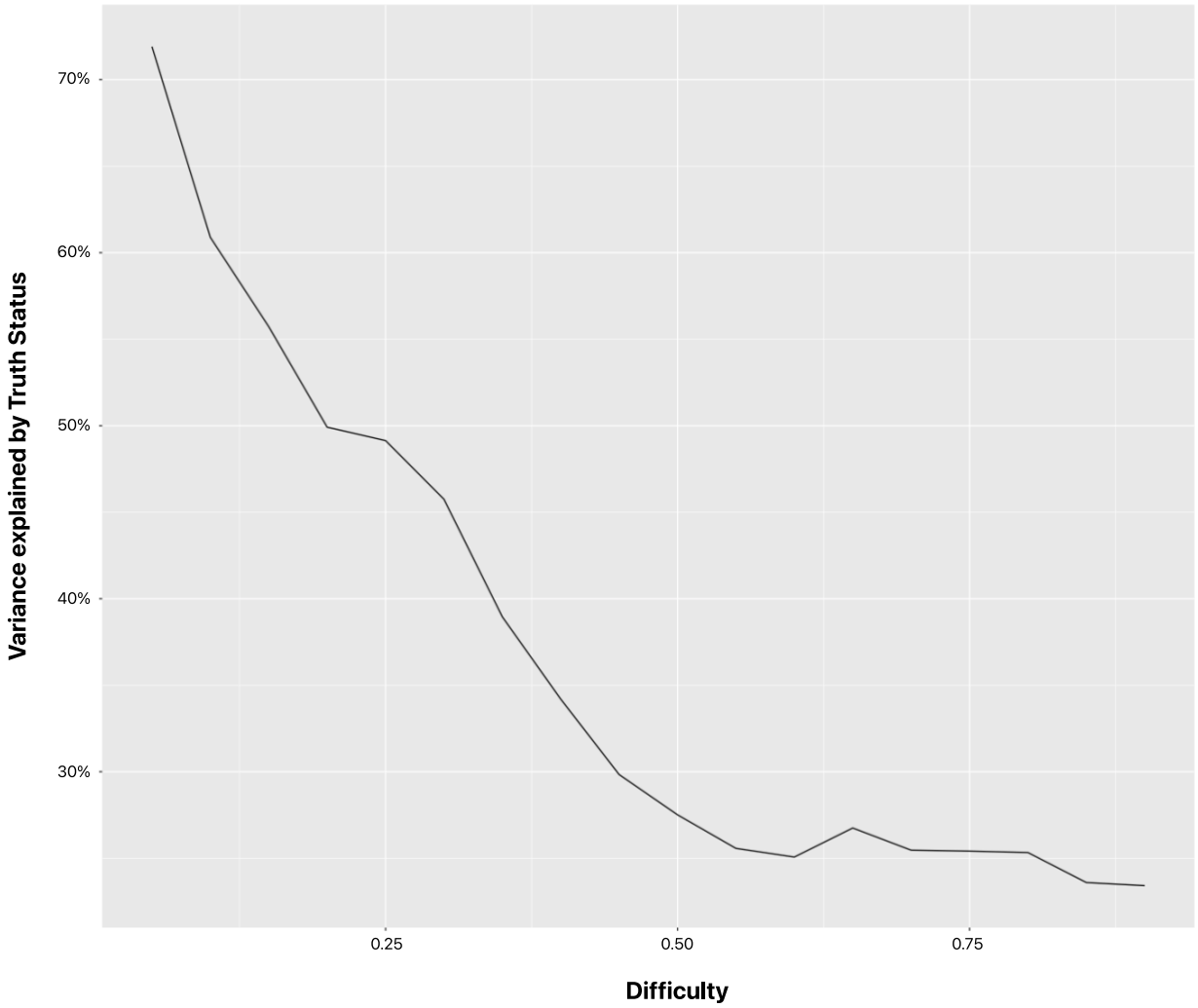
	B	95% CI		β	*p*-value
		LL	*UL*		
Intercept (Hit)	4.87	4.45	5.3	0	< .001 *
Recollection Success Miss	-1.15	-2.3	0	-.16	.05
Repetition	-.45	-.58	-.32	-.1	< .001 *
Interaction Recollection Success Miss x Repetition	.5	.07	.93	.19	.022 *

Note. This table shows predictors for Model 5. The regression coefficient is shown by *B*, 95% confidence intervals are shown including the lower limit (*LL*) and the upper limit (*UL*). Furthermore, standardized beta-coefficient is indicated by *β*. The intercept represents statements were correctly identified during the retrieval practice. Significance level of *p*-values are *p < .05

Appendix C

Figure C1

Variance Explained by Truth Status as a Function of Difficulty



Note. This plot shows the variance in the truth rating variable explained by the truth status variable as a function of statement difficulty. For easier questions, truth status was able to explain most of the variance in the truth rating variable. With increasing difficulty, the explanatory power of truth status decreased.