14 Illusory truth effect

Lena Nadarevic

It is common knowledge that repeated studying leads to better memory than a single study episode. But repetition not only strengthens memory for information, it also increases the judged truth of information (Hasher et al., 1977). This truth effect emerges irrespective of whether the processed information is true or false. For this reason, the effect is also referred to as illusory truth effect. Further synonyms in the literature are validity effect, reiteration effect, and truth-by-repetition effect.

Relevance of the truth effect

While most research on the truth effect has been conducted under controlled experimental conditions and with student samples, several studies show that the effect also replicates in field experiments (Boehm, 1994), with representative samples (Gigerenzer, 1984), and with naturalistic stimuli such as social media postings (Pennycook et al., 2018). First empirical work on the truth effect was published several decades ago (Hasher et al., 1977), yet interest in the effect has not diminished. To the contrary, interest in the truth effect is actually increasing, particularly with regard to its role in people’s belief in misinformation and fake news (see Chapter 20). Indeed, recent studies showed that repetition enhances the perceived accuracy of true and fake news headlines (Pennycook et al., 2018; Smelter & Calvillo, 2020). Text box 14.1 provides further examples of the truth effect in everyday contexts.

Text box 14.1 Real-world examples of the truth effect

- Coming across a news headline repeatedly on social media increases the perceived accuracy of the headline, even if it is discordant with one’s own political view (Pennycook et al., 2018).
- Repetition also boosts people’s belief in advertising claims (Hawkins & Hoch, 1992) and subjective opinions (Arkes et al., 1989).
- Repeated eyewitness testimony is perceived as more credible, even if the repeated testimony stems from a single source (Foster et al., 2012).
- Items presented in true/false tests and multiple-choice tests are judged more likely as true, if they are familiar from a previous test. This holds for correct items as well as distractors, at least in the absence of test feedback (Toppino & Luipersbeck, 1993).
Experimental designs and measures

Classical paradigm

The classical research paradigm on the truth effect is based on the first empirical truth-effect study by Hasher et al. (1977) and consists of at least two phases. In Phase 1, participants are presented with a list of statements. The typical stimuli used in truth-effect studies are trivia statements, half of which are factually true (e.g., Manama is the capital of Bahrain) and half of which are factually false (e.g., Bolivia is the smallest landlocked country of South America). Ideally, these statements are pretested and matched on plausibility. The participants’ task in the first phase is to rate the truth of the presented statements on a Likert scale ranging from definitely false to definitely true or by providing binary true/false judgments. After several days or even weeks, participants take part in Phase 2. Again, the task is to judge the truth of several statements. This time, however, some of the statements are repeated statements from Phase 1 while the others are new.

In the classical design, there are two ways to investigate the truth effect (Dechêne et al., 2010). To test for a within-items truth effect, one exclusively focuses on the truth judgments for the repeated statements, that is, the statements presented in both judgment phases. A within-items truth effect is present if the truth judgments for these repeated statements (dark bars in Figure 14.1) are significantly higher in Phase 2 (i.e., at repeated exposure) compared to Phase 1 (i.e., at first exposure). In contrast, to test for a between-items truth effect, one exclusively focuses on the truth judgments in Phase 2. A between-items truth effect is present if the truth judgments for the repeated Phase 2 statements are significantly higher than the truth judgments for the new Phase 2 statements (see Figure 14.1).

![Figure 14.1](image.png)

**Figure 14.1** The truth effect in the classical paradigm.

*Note:* Mean truth ratings (1 = definitely false; 7 = definitely true) reported by Hasher et al. (1977). For ease and clarity, the graph covers only two judgment phases and the y-axis is truncated.
**Exposure paradigm**

Because the truth effect does not replicate in the classical paradigm if the retention interval between the two judgment phases is only a few minutes instead of several days (Nadarevic & Erdfelder, 2014), many truth-effect studies use a slightly adapted procedure – the exposure paradigm. This paradigm is more economical because it allows researchers to examine the truth effect within a single experimental session. As in the classical paradigm, participants are exposed to a list of statements in the first phase of the experiment. In this exposure phase, however, participants do not judge the truth of the statements. Instead, they assign the statements to different semantic categories (Nadarevic & Erdfelder, 2014), judge the interestingness of the statements (Fazio et al., 2015), simply read the statements (Unkelbach & Rom, 2017), or engage in other processing tasks. The second phase, often referred to as the test phase, typically takes place directly after the exposure phase or shortly after. As in the classical paradigm, participants are asked to judge the truth of statements, some of which are repeated statements from the exposure phase while the others are new.

Because participants only provide truth judgments in the test phase, it is not possible to test for a within-items truth effect in the exposure paradigm. Therefore, it is even more important that researchers take great care in pretesting and counterbalancing their stimuli to ensure that a between-items truth effect cannot be attributed to factors other than statement repetition. Despite this drawback, the exposure paradigm has several advantages over the classical design. First, as noted above, it is much easier to implement (for a classroom demonstration, see Text box 14.2). Second, the exposure design is characterized by higher external validity than the classical design. People typically do not pay attention to accuracy when they read or listen to information. Thus, the exposure phase more closely resembles how participants process new information in everyday life.

**Memory paradigm**

Finally, there are truth-effect studies in which the participants receive cues or information on the truth value of the presented statements in the first phase of the experiment – the study phase. Except for this credibility information in Phase 1, the procedure of this memory design is comparable to the other truth-effect paradigms, that is, participants provide truth judgments on repeated and new statements in a later test phase. In case of binary truth judgments, the following pattern of results indicates a truth effect in the memory design: First, the proportion of repeated, false statements judged as true is larger than the proportion of repeated, true statements judged as false. Second, the proportion of repeated, false statements judged as true is larger than the proportion of new statements judged as true. In the case of truth ratings, the second criterion (i.e., higher truth judgments for repeated, false statements compared to new statements) is typically adopted. Because the truth effect in the memory paradigm largely depends on people’s memory for the truth-value information provided in the study phase, it is no longer a judgment effect in the strict sense. For this reason, this chapter focuses on the truth effect in the classical design and the exposure design.
Text box 14.2 Classroom demonstration of the truth effect

Participants

Ideally, there should be at least 27 participants in the class. This group size allows to detect a medium-sized truth effect (Cohen’s $d_z = 0.50$) with a power of at least .80, when analyzing the data with a one-tailed $t$-test ($\alpha = .05$).

Materials

A minimum number of 20 statements is required. These statements are assigned to two stimulus sets matched on plausibility. Ideally, the sets should contain an equal number of true and false statements (two sample sets are provided in the Appendix).

Procedure

Students are randomly assigned to two groups (e.g., A and B). Group A receives a sheet with all Set-A statements and Group B a sheet with all Set-B statements listed in random order. The students’ task is to rate the interestingness of the statements (e.g., on a six-point scale ranging from not interesting at all to very interesting). After this exposure phase, the first sheet is replaced by a second one, which contains the statements of both sets listed in random order. This second sheet is essentially the same for both groups. However, the correct group name (A or B) needs to be listed on the sheet in order to allow the later coding of repeated and new statements. This time, the students’ task is to rate the truth of the presented statements (e.g., on a six-point rating scale ranging from definitely false to definitely true).

Statistical analysis

Depending on group membership, statements from one set are coded as repeated and statements from the respective other set as new. Mean truth judgments are then compared between the repeated and the new statements. Higher truth judgments for the repeated statements indicate the expected between-items truth effect. The statistical significance of this effect can be tested with a one-tailed $t$-test for dependent means.

Theoretical accounts

The following sections cover the most popular explanations of the truth effect.

Familiarity account

It has been proposed that the credibility-enhancing effect of repetition is mediated by familiarity (e.g., Arkes et al., 1991; Boehm, 1994). Support for this familiarity account comes from truth-effect studies in which participants judged not only the truth but also the familiarity of the presented statements. For instance, Boehm (1994) found that repetition enhanced familiarity, and familiarity enhanced judged truth. Moreover, the direct
effect of repetition on judged truth disappeared when controlling for familiarity. But how to explain the link between familiarity and judged truth?

Arkes et al. (1991) argued that when people experience a high familiarity with a given statement, they tend to misattribute this familiarity to a pre-experimental exposure to the statement. Moreover, the authors reasoned that such source dissociations increase judged truth, because people infer that they have heard or seen these statements from two independent sources (i.e., an external source and the experiment). Indeed, statements attributed to external sources receive considerably higher truth ratings than statements attributed to the experiment or judged to be new (Arkes et al., 1989). However, although source dissociations seem to contribute to the truth effect, the effect persists even when excluding truth judgments for statements attributed to external sources.

Alternatively, it has been proposed that people rely on familiarity when judging truth because it is an ecologically valid cue for truth (Reber & Unkelbach, 2010). In fact, if most people adhered to the maxim of quality (Grice, 1989) in their everyday communications, familiar true information should clearly outnumber familiar false information. Moreover, Unkelbach and Stahl (2009) reasoned that “there are countless possible false propositions about properties of the physical world, but only one true proposition” (p. 23). For this reason, the probability of repeatedly encountering a true statement (e.g., Manama is the capital of Bahrain) should be higher than encountering any related false statement (e.g., Manama is the capital of Qatar). Note, however, that this argument does not hold for false statements that are repeated and spread on purpose (e.g., advertisements, political propaganda, or fake news).

Fluency account

The term fluency denotes the subjective, metacognitive experience of processing ease. Under certain circumstances (which will be specified below), this experience elicits a feeling of familiarity (e.g., Whittlesea & Williams, 2000). Hence, if familiarity drives the truth effect, it should be possible to induce the effect by means of perceptual fluency manipulations instead of repetition. Indeed, Reber and Schwarz (1999) could replicate the truth effect by manipulating the readability of the presented statements. Statements that appeared in high-contrast colors on a white background, which made them easy to read, were characterized by higher truth ratings than statements that appeared in low-contrast colors on the white background, which made them more difficult to read. Other perceptual fluency manipulations produced similar results (e.g., Parks & Toth, 2006).

According to the discrepancy-attribution hypothesis (Whittlesea & Williams, 2000), fluency only elicits a feeling of familiarity if the stimulus is processed surprisingly fluently, that is, if there is a discrepancy between the expected and the experienced fluency. Several studies support this hypothesis (see Wänke & Hansen, 2015, for a review). For example, Dechêne et al. (2009) did not find a truth effect when manipulating statement repetition between subjects, with one group judging only repeated statements and another one judging only new statements in Phase 2. The authors reasoned that, unlike in mixed lists of new and repeated statements, people do not experience any fluency discrepancy in homogeneous lists. Although Garcia-Marques et al. (2019) did find a truth effect in a between-subjects design, the effect diminished significantly over the course of the statement list. This finding suggests that participants dynamically adjust their fluency expectations.

Even though perceptual fluency manipulations can induce illusory truth, perceptual truth effects are typically much smaller and less robust than the repetition-based
truth effect (e.g., Parks & Toth, 2006; Silva et al., 2016). Recently, Vogel et al. (2020) hypothesized that fluency effects depend on the match between the type of fluency (perceptual versus conceptual) and the judgment task (perception-related versus content-related). The authors tested this fluency-specificity hypothesis by orthogonally manipulating the perceptual fluency (by color contrast) and conceptual fluency (by content repetition) of statements. As predicted, conceptual fluency had a stronger effect than perceptual fluency on judgments of truth whereas the reverse pattern emerged for judgments of aesthetic pleasure. This finding is also in line with the assumption of Schwarz (2004) that people hold naïve beliefs about their metacognitive experiences. These naïve beliefs may influence whether or to what degree a certain experience is considered informative for the judgment at hand.

**Referential theory**

The referential theory of the truth effect (Unkelbach & Rom, 2017) proposes that truth judgments depend on the activation and coherence of localized networks in people’s semantic memory. Specifically, the theory proposes that, when reading a statement, memory references get activated and linked. That is, building on the idea of spreading-activation models, existing links between the references are strengthened and new links start to form, which results in a localized information network. Thus, according to the theory, repeated statements are perceived as more likely true because they are characterized by more coherently linked references than novel statements. Conversely, incoherent links between references should increase the likelihood of judging the statement as “false”. Accordingly, statements that contradict previously processed statements should be perceived a less likely true than new statements. In fact, such an illusion of falseness has been found in several studies, at least in case of a short retention interval between the initial statement presentation and its contradicting repetition (García-Marques et al., 2015; Nadarevic et al., 2020; Silva et al., 2017; Unkelbach & Rom, 2017).

Importantly, Unkelbach and Rom (2017) also addressed the role of processing fluency in their referential theory. They argued that coherent statement processing produces a feeling of fluency. However, this fluency experience is considered as an output variable and not as the central mediator between repetition and perceived truth in their theory. Yet, because fluency typically accompanies perceptions of truth, the theory proposes that people learn to associate fluency with truth. This association can lead to illusions of truth “when fluency and truth are factually orthogonal, or when fluency is manipulated independent of repetition” (Unkelbach & Rom, 2017, p. 113).

**Integrative model**

Unlike competing theories concerning many other phenomena, the presented theories on the truth effect should not be considered as competitors. That is, the proposed explanatory approaches are not mutually exclusive, but rather complement each other. For example, fluency-based explanations supplement familiarity-based explanations by making clear predictions about the circumstances under which fluency is interpreted as familiarity. The referential theory, on the other hand, provides a process-based explanation for the source of conceptual fluency. Moreover, it can also account for the significantly smaller effect of perceptual fluency on judgments of truth. Given the high interrelatedness of the cognitive constructs that have been proposed to underlie the truth effect
(i.e., familiarity, fluency, semantic coherence), Unkelbach et al. (2019) proposed an integrative model. This model assumes that all of the above-mentioned constructs may directly or indirectly contribute to the truth effect. Possibly, future research will enable us to disentangle and to estimate the contribution of these different causes of the truth effect.

Moderators of the truth effect

As outlined in the introduction, the truth effect is a very robust effect that occurs with different materials, in different contexts, and within different sample populations. There are only very few studies that have identified boundary conditions of the effect. The truth effect does not replicate with extremely implausible statements (e.g., the earth is a perfect square, Pennycook et al., 2018), with a short retention interval in the classical paradigm (Brashier et al., 2020; Calvillo & Smelter, 2020; Nadarevic & Erdfelder, 2014), and with a homogeneous list of repeated statements (Dechêne et al., 2009; but see Garcia-Marques et al., 2019). With those exceptions aside, the truth effect has been found in more than 100 studies (Fazio & Sherry, 2020). A meta-analysis of Dechêne et al. (2010) reported an average effect size of $d = .49$, 95% CI [.45, .55] for the between-items truth effect and $d = .39$, 95% CI [.32, .47] for the within-items truth effect. However, several variables have been identified that moderate the size of the effect. The following sections give an overview on different a) statement characteristics, b) context characteristics, and c) person characteristics that have been examined as potential moderators of the effect.

Statement characteristics

As becomes evident from the examples presented in Text box 14.1, the truth effect has been observed for a variety of statement types. Moreover, the effect does not depend on the presentation time of statements and appears equally strong for visually and auditorily presented statements (Dechêne et al., 2010). In their meta-analysis, Dechêne et al. (2010) also compared the effect of verbatim statement repetition and gist repetition (i.e., reiteration of meaning, but not of wording) and found a larger between-items truth effect for verbatim repeated statements. For the within-items truth effect, in contrast, the number of studies with gist repetition was too small to perform a moderator analysis. Silva et al. (2017) conducted a direct test of verbatim versus gist repetition on the between-items truth effect. Their experiments did not reveal any significant differences between the two. These findings underline the importance of conceptual rather than perceptual fluency for the truth effect.

Some findings suggest that the truth effect only replicates if the presented statements stem from domains about which people are at least moderately knowledgeable (Arkes et al., 1989; Boehm, 1994). In contrast, Unkelbach and Rom (2017) observed a truth effect even for completely meaningless statements (e.g., A ma is bigger than an omp), although the effect was smaller than for meaningful statements. While Arkes et al. (1989) and Boehm (1994) investigated prior knowledge about the statement domains, they held the likelihood of prior knowledge about individual statements constant. In fact, early truth-effect studies had exclusively used statements for which knowledge about the correct truth status was very unlikely (so-called difficult statements). This choice of material rested on the widespread assumption that the truth effect could only occur for difficult statements. Fazio et al. (2015) were the first to put this assumption to the test.
At odds with the aforementioned assumption, their experiments disclosed that the truth effect also replicates for easy statements, that is, statements that are easily identifiable as true or false according to pretest norms. Similarly, participants even showed a truth effect for statements for which they demonstrated relevant knowledge in a later knowledge test. Furthermore, a simulation study by Fazio et al. (2019) suggests that, in principle, repetition boosts the perceived truth for easy and difficult statements alike. Empirically, however, ceiling effects counteract the truth effect for extremely plausible statements. Likewise, the midpoint of the truth judgments scale represents a ceiling for extremely implausible statements because people will refrain from judging these statements as true, despite an internal increase of perceived truth.

Nadarevic et al. (2018) investigated whether statement language moderates the truth effect. The authors presumed that the activation of semantic network references might be weaker when processing foreign-language statements as compared to native-language statements. This should lead to a smaller truth effect according to the referential theory. However, Nadarevic et al. (2018) did not observe any differences in the truth effect between a foreign-language and a native-language group, at least when the test phase followed the exposure phase in close succession. In contrast, after a two-week retention interval, the truth effect was significantly smaller in the foreign-language group. Overall, this pattern of results is incompatible with the authors’ activation hypothesis. Instead, it indicates a faster decay of semantic networks in a foreign language. Although further research on foreign-language effects on truth judgments is still warranted, the findings by Nadarevic et al. (2018) show that contextual variables, such as the length of the retention interval, can play a crucial role in truth-effect studies.

**Context characteristics**

In fact, the length of retention interval between initial statement exposure and repeated statement exposure varies strongly between different truth-effect studies. In their meta-analysis, Dechène et al. (2010) did not find a significant effect of repetition delay on both, the between-items and the within-items truth effect. However, the results of later studies suggest that the influence of retention interval on the truth effect depends on the experimental paradigm. In the exposure paradigm, the truth effect tends to be larger when initial statement exposure and repeated exposure take place within the same experimental session than when they are separated by a one-week delay (e.g., Silva et al., 2017; Stump et al., 2021). Conversely, in the classical design the effect does not appear within one session but after a week (Nadarevic & Erdfelder, 2014).

Most studies on the truth effect have compared truth judgments for once-repeated statements and new statements. Few studies (e.g., Gigerenzer, 1984; Hasher et al., 1977; Pennycook et al., 2018) also implemented a second repetition and found a larger truth effect for twice-repeated than for once-repeated statements. However, the increase of the truth effect due to the second repetition was rather small (but see Pennycook et al., 2018). Going further, Hawkins et al. (2001) included up to four repetitions and DiFonzo et al. (2015) up to nine repetitions in their experiments. Both studies found a logarithmic relationship between number of repetitions and subjective truth ratings. But not all researchers observed an increase of the truth effect with further repetitions (see Arkes et al., 1991). Fazio et al. (2021) hypothesized that the size of the truth effect not only depends on the number of repetitions, but also on the spacing of repetitions. In line with prior studies, the authors found a logarithmic increase of the truth effect across repetitions. But, unlike
what was predicted, this pattern was not moderated by spacing (one day versus four days between each repetition).

Specific to the exposure paradigm it has also been studied whether the truth effect depends on the processing task in the exposure phase. For example, in a study by Hawkins and Hoch (1992), participants were either instructed to rehearse the statements in the exposure phase (rote rehearsal task), to judge the comprehensibleness of the statements (comprehension task), or to count the number of a specific letter within a statement (orthographic task). The rote rehearsal group showed the largest truth effect followed by the comprehension group. In the orthographic group, in contrast, the effect was not significant. Based on this finding the authors concluded that “there appears to be a minimum level of processing that must occur for the truth effect to take place” (p. 222). Overall, the truth effect seems to be larger when the processing task increases memory for the content of the statements (see also Unkelbach & Rom, 2017).

What is less clear is the role of people’s processing capacity when judging truth. Two research papers on this topic obtained different results. Garcia-Marques et al. (2016) manipulated participants’ processing capacity in the test phase by means of cognitive load. They found a smaller truth effect in a low-load condition compared to a high-load condition, at least when task instructions emphasized accurate judgments. When the instructions emphasized intuitive judgments, in contrast, the truth effect was not affected by cognitive load. Nadarevic et al. (2021) investigated the truth effect under time-pressure conditions. Participants either had to provide very fast truth judgments in the test phase or could take as much time as they wanted for the truth judgments. Unlike cognitive load, time-pressure did not moderate the truth effect, and this null effect persisted with different task instructions and different response deadlines in the time-pressure group.

**Person characteristics**

Typically, the truth effect has been studied with student samples of young adults. Thus, it is an interesting question whether the effect is stable across the lifespan. Studies that compared the effect between young and old adults produced mixed results. On a descriptive level, the meta-analysis of Dechêne et al. (2010) speaks in favor of a larger truth effect for older adults aged about 73 years \( (d = 0.64) \) compared to younger adults aged about 23 years \( (d = 0.49) \). However, this difference did not reach statistical significance. Surprisingly, there has been only a single study to date that has examined the truth effect in children (Fazio & Sherry, 2020). This study, which examined the truth effect in 5-year-olds, 10-year-olds, and a group of adults, showed that the effect is already present by the age of 5. What is more, the truth effect did not differ significantly between age groups. However, because Fazio and Sherry (2020) had not designed their study to detect effect-size differences between age groups, this null effect must be interpreted with caution.

A number of studies aimed at exploring whether individuals with certain personality traits are particularly susceptible to the truth effect. Arkes et al. (1991) and Boehm (1994) were the first to investigate interindividual differences in the truth effect. The authors tested a possible relationship between the truth effect and participants’ Need for Cognition (NfC, Cacioppo & Petty, 1982), but did not detect a relationship. Newman et al. (2020), in contrast, found tentative evidence that participants high in NfC show a larger truth effect than participants low in NfC. However, this effect was only evident when participants were not informed that they would see true and false statements in the exposure phase. De Keersmaecker et al. (2020) explored whether any of the following
constructs accounts for individual differences in the truth effect, but failed to find an association: cognitive ability (e.g., verbal intelligence), cognitive style (e.g., preference for intuition or deliberation, respectively), and Need for Cognitive Closure (NCC, Webster & Kruglanski, 1994). In contrast, a recent study by Stump et al. (2021) found a larger truth effect for people high in NCC than for people low in NCC. In one of their experiments, however, this relationship was evident only after a ten-minute retention interval but not after a one-week interval.

Taken together, procedural differences such as differences in task instructions (Newman et al., 2020) or differences in the length of retention interval (Stump et al., 2021) could account for the inconsistent findings on interindividual correlates of the truth effect. However, methodological differences could also play an important role. One general problem is that mean-difference scores of the truth effect ($M_{repeated} - M_{new}$) are unreliable at the individual level (test-retest reliability: $r \leq .12$, Calio et al., 2021). Based on this finding, it is not particularly surprising that studies that have used mean-difference scores to examine correlates of the truth effect (e.g., De Keersmaecker et al., 2020) failed to find reliable associations. In contrast, it is more reasonable to make use of regression models or mixed linear models to test whether certain trait variables moderate the effect of statement repetition on judged truth (e.g., Newman et al., 2020; Stump et al., 2021). What is more, using a Bayesian modeling approach, Schnuerch et al. (2020) showed that people not only differ in the strength of the truth effect, but that some people even show a reversed truth effect where novel statements receive higher truth judgments than repeated ones. But it is still an open question how stable such reversals are and what causes them.

**The truth effect under naturalistic conditions**

Even though the truth effect is an extremely robust phenomenon on the group level, one might ask whether it has any relevance under naturalistic conditions. In the meta-analysis by Dechêne et al. (2010), the size of the effect was moderate. But how strong is the influence of repetition on truth judgments in the real world? A study by Jalbert et al. (2020) suggests that the truth effect tends to be considerably larger in naturalistic settings than in most laboratory studies. This is because in truth-effect experiments, participants are usually “warned” before the exposure phase that they will see true and false statements. In many studies, participants are even informed about the actual base rates of true and false statements. In contrast, in the real world, there are typically no prior warnings that precede information processing. Without typical warning instructions, the size of the truth effect increased to $d = 1.55$, 95% CI [1.33, 1.76] in the study by Jalbert et al. (2020), which is three times larger than the effect size reported by Dechêne et al. (2010).

Most studies of the truth effect have used minimalistic designs to investigate the isolated effect of statement repetition on judgments of truth. Under these conditions, it is not particularly surprising that participants rely on metacognitive feelings (e.g., familiarity or fluency) when judging truth, because there are no other judgment cues available in the given context. But is the effect of repetition still present in a more complex environment (e.g., in a social-media context) that includes other cues? Nadarevic et al. (2020) investigated this question as follows: Participants were presented with repeated and new statements that mimicked social media news postings. Each statement appeared either with a source high in credibility (e.g., a real, trustworthy news source), a source low in credibility (e.g., a fabricated news source), or without any source information. In addition,
Each statement was either accompanied by a thematically related but non-probative picture (as is often the case with social media news) or appeared without a picture. Under these conditions, Nadarevic et al. (2020) found significant effects of statement repetition and source information on judged truth, but no effect of non-probative pictures. Importantly, even though participants relied on two different cues for truth (repetition and source information), they used these cues in an additive fashion.

It is also important to note that the truth effect is not limited to individual statements or headlines, but also replicates for short (false) news stories. In a study by Polage (2012), participants provided higher plausibility and truth ratings to false news stories when they had been previously exposed to these stories (exposure group) than when they had not seen the stories before (control group). Moreover, participants in the exposure group rated themselves to be more knowledgeable about the covered topics and were more prone to source misattribution errors. That is, they were more likely to believe that they had previously heard the stories from a source outside of the experiment.

Debiasing approaches

As the truth effect contributes to people’s belief in misinformation, political propaganda, and false advertising promises, the question is how to eliminate or at least reduce this effect. Several debiasing approaches have been tested so far with different levels of success. The first debiasing study on the truth effect was conducted by Boehm (1994), who told one group of participants that they would have to justify their truth judgments to a group of peers. However, this accountability manipulation did not have any effect on the size of the truth effect. Nadarevic and Abfalg (2017) tested whether it is possible to eliminate the truth effect by warning people about it. In their study, one group of participants received a warning before the test phase that informed them about the truth effect and instructed them not to show the effect. Even though the warning did not eliminate the truth effect, the effect was at least smaller for warned than for unwarned participants. A follow-up study by Calio et al. (2020) produced similar findings. Moreover, in a separate analysis of easy and difficult statements, the authors observed an elimination of the truth effect for easy statements in the warning group.

One debiasing approach that has been implemented by social-media platforms to combat misinformation is the tagging of misleading or false information with warning labels or corrections. Pennycook et al. (2018) tested the effectiveness of this intervention in a truth-effect experiment that involved fake and real news headlines. During the exposure phase, fake-news headlines appeared with a warning label (“Disputed by 3rd Party Fact-Checkers”) in a warning group and without any label in a no-warning group. Even though the warning label reduced participants’ willingness to share the tagged headlines on social media, the tagging did not prevent a truth effect for repeated fake-news headlines. That is, both, the warning group and the no-warning group showed a truth effect in a subsequent test phase, although the effect was somewhat smaller in the warning group.

Some studies speak to the crucial role of initial information encoding in the truth effect. For instance, in the exposure paradigm, the effect reduces drastically when participants are informed prior to information encoding that they will be exposed to true and false information (Jalbert et al., 2020). In the classical design, in which individuals rate the truthfulness of statements at the first encounter, the effect even disappears completely, at least in case of a short retention interval (Calvillo & Smelter, 2020; Nadarevic & Erdfelder, 2014).
Brashier et al. (2020) showed that focusing on the accuracy of information at exposure still prevented a truth effect two days later. However, this was only the case for statements for which the participants held prior knowledge.

Finally, research by Corneille et al. (2020) suggests that focusing on the fakeness rather than the truth of information may be an effective means to counter the truth effect for social-media misinformation. When the authors instructed participants to identify statements as fake news instead of judging their truth, participants even categorized repeated statements more likely as fake news than new statements. This finding suggests that in principle people are able to reinterpret familiarity (or fluency, respectively) in the context of social media. Hence, the mental reappraisal of metacognitive feelings could be an effective means to eliminate the truth effect. However, more applied studies on this approach are still pending.

Conclusion

The truth effect is a very robust judgment bias. In contexts where true information clearly predominates (e.g., educational settings), the effect is quite adaptive, largely leading to correct truth judgments. However, in contexts where people are frequently exposed to misinformation (e.g., advertising, social media), the effect contributes to people’s belief in falsehoods. Thus, in the latter context, individuals are well-advised to avoid processing information from untrustworthy or unknown sources as there is currently no intervention that reliably prevents a truth effect for misinformation. What is more, effects of misinformation on people’s beliefs are difficult to correct once the false information has been encoded (Ecker et al., 2011). The authors of the Debunking handbook 2020 (Lewandowsky et al., 2020) advise, among other things, that refutations are best presented multiple times to counter misinformation. Indeed, it makes perfect sense to fight the truth effect with its own weapon—repetition.

Summary

- People judge repeatedly presented statements to be more likely true than new statements—a phenomenon referred to as (illusory) truth effect.
- A typical truth–effect experiment consists of two phases. In Phase 1, participants read or listen to statements which they either judge for truth (classical paradigm) or process in some other way (exposure paradigm). In Phase 2, participants are asked to judge the truth of new statements as well as of repeated statements from Phase 1.
- Higher truth judgments for the repeated Phase 2 statements than for the new Phase 2 statements indicate a between-items truth effect. The classical paradigm also allows to test for a within-items truth effect, that is, an increase in truth judgments in Phase 2 compared to Phase 1 for the repeated statements.
- Different assumptions have been made about which variables mediate the repetition effect on perceived truth, such as familiarity, fluency, and semantic coherence. Given the high relatedness of these constructs, an integrative model encompassing all of these variables currently seems most appropriate to explain the truth effect.
- Although the truth effect is a quite robust phenomenon, its effect size tends to vary depending on certain characteristics of the statements, the context, and the person. New findings suggest that the truth effect not only replicates under more naturalistic conditions, but is likely to be even larger under these conditions.
• To counteract the truth effect for false information, several debiasing approaches have been proposed and tested. Although none of these approaches has reliably prevented a truth effect, the suggested interventions could be quite effective when applied in combination.

Further reading
Hasher et al.'s (1977) first article on the truth effect is a good read to get into the topic. The only meta-analysis of the truth effect so far is provided by Dechêne et al. (2010). It covers important findings on the effect published before the year 2010. A more timely theoretical review article stems from Unkelbach et al. (2019). Going beyond the truth effect, a review by Brashier and Marsh (2020) provides a systematic overview of various determinants of truth judgments.

Acknowledgment
I am grateful to Martin Schnürch and Alina Kias for their valuable comments on an earlier version of this chapter.

References


Illusory truth effect

Vogel, T., Silva, R. R., Thomas, A., & Wänke, M. (2020). Truth is in the mind, but beauty is in the eye: Fluency effects are moderated by a match between fluency source and judgment dimension. *Journal of Experimental Psychology: General, 149*(8), 1587–1596.

**APPENDIX**

Trivia statements for a classroom demonstration taken from prior truth-effect studies (Nadarevic & Aßfalg, 2017; Nadarevic & Erdfelder, 2014).

**Set A**
- The area of skin between the eyebrows is called “glabella”. (true)
- Manama is the capital of Bahrain. (true)
• The game of dominoes is played with 28 tiles. (true)
• March and April are the hottest months in the Maldives. (true)
• Italy won the football world championship for the first time in 1934. (true)
• A dingo can survive longer without water than a camel. (false)
• Ultrasound is sound waves with frequencies below 16 Hertz. (false)
• The all-wheel drive was invented in England in 1938. (false)
• Sumatra is the fourth biggest island of the world. (false)
• The Montgolfier brothers created the first parachute. (false)

Set B
• There are three American cities named “Santa Claus”. (true)
• Mongols flavor their tea with salt instead of sugar. (true)
• “Ligne” is a unit of length used to measure the diameter of buttons. (true)
• The Ad-Dahna desert is located in Saudi Arabia. (true)
• Canada has the longest coastline of any nation worldwide. (true)
• Hoverflies are the only insects that are able to fly backwards. (false)
• Bolivia is the smallest landlocked country of South America. (false)
• The kilt originated in Ireland and not in Scotland. (false)
• In the film *Pulp Fiction* all clocks are set to 4:10. (false)
• The world’s highest giant mammoth tree is called “King Arthur”. (false)