Plausibility matters: A challenge to Gilbert's "Spinozan" account of belief formation

Marion Vorms*

University Paris 1 Panthéon-Sorbonne, IHPST

Adam J. L. Harris and Sabine Topf

University College London

Ulrike Hahn

Birkbeck College, London

*Corresponding author: mvorms@gmail.com

Declaration of interests: none

Abstract

Most of the claims we encounter in real life can be assigned some degree of plausibility, even if they are new to us. On Gilbert's (1991) influential account of belief formation, whereby understanding a sentence implies representing it as true, all new propositions are initially accepted, before any assessment of their veracity. As a result, plausibility cannot have any role in initial belief formation on this account. In order to isolate belief formation experimentally, Gilbert et al. (1990) employed a dual-task design: if a secondary task disrupts participants' evaluation of novel claims presented to them, then the initial encoding should be all there is, and if that initial encoding consistently renders claims 'true' (even where participants were told in the learning phase that the claims they had seen were false), then Gilbert's account is confirmed. In this pre-registered study, we replicate one of Gilbert et al.'s (1990) seminal studies ("The Hopi Language Experiment") while additionally introducing a plausibility variable. Our results show that Gilbert's 'truth bias' does not hold for implausible statements — instead, initial encoding seemingly renders implausible statements 'false'. As alternative explanations of this finding that would be compatible with Gilbert's account can be ruled out, it questions Gilbert's account.

Keywords

Belief formation; Truth bias; Plausibility; Credulity.

Plausibility matters: A challenge to Gilbert's "Spinozan" account of belief formation

When encountering real-world claims, it is typically possible to assign some degree of plausibility to them, even when such claims are entirely new. For example, you may not know, nor have ever wondered, what the speed limit on expressways in New Caledonia is, but you would probably find the claim that it is 350 mph rather implausible (and you would be right!). What role does plausibility play in agents' belief states for new propositions?

According to Gilbert's (1991) so-called 'Spinozan' theory of belief formation, any new proposition is initially encoded as true. Only then is it assessed and, possibly, re-tagged as false. Indeed, Gilbert claims that acceptance of a proposition is part of its very comprehension: understanding a proposition implies representing it as true. Consequently, it is not possible to suspend belief before evaluating veracity: people have no choice but to first believe any new statement they read or hear. Only in a second stage may they then assess its veracity and decide whether to keep believing it or reject it. On this account, plausibility can have no role in initial belief formation.

Gilbert's theory is more than an esoteric idea about how the human cognitive system functions. There are empirical results to support it (Gilbert et al., 1990, 1993), and it is of both theoretical and practical consequence in several areas. Through an ingenious experimental design (described below), Gilbert et al. (1990) tackle the question of the very *formation* of belief — of how people accept entirely novel contents. The implications of such an original perspective for the very nature of belief are likely among the reasons why their work has a reception far beyond the bounds of psychology. For instance, Mandelbaum (2013) takes Gilbert's theory as a basis for a philosophical account of mental states. Gilbert's work is also often cited as a potential explanation for the so-called 'confirmation bias' (see e.g., Nickerson, 1998) and has been taken to support dual system theory (see Kahneman, 2003). Perhaps most urgently, it has gained renewed relevance in the context of the contemporary inquiry into 'fake news' and the human psychology of dealing with false claims and their subsequent correction (see e.g., Lewandowsky et al., 2012).

Re-examining Gilbert's high-profile work seems both timely and relevant. Specifically, Gilbert's claims about default encoding appear at odds with the very notion of 'plausibility', as highlighted by our New Caledonian traffic law example. Would one not expect memory encoding to be sensitive to degrees of plausibility? If Gilbert's theory holds, any statement, however implausible it might be, should initially be believed. As our opening example demonstrates, this seems somewhat unintuitive.

It makes sense to assume that comprehending a claim involves understanding the conditions under which it would be true (and much philosophy of language has involved explicating this intuition, see Davidson, 1967; Dummett, 1959, 1976; Higginbotham, 1986, 1989). Yet it is unclear why such understanding could not be entirely hypothetical: comprehending that '*the moon is made of cheese*' may require understanding what the moon would (hypothetically) have to be like were that statement to be true. But why would the cognitive system have to assume that these conditions actually hold for comprehension to proceed? The same goes for a statement preceded by the information that it is a lie: *'The president told a lie today when he said that…'.* It is thus important to examine the role of plausibility in initial belief formation.

Crucial to this endeavour is pinning down 'initial belief' because Gilbert's account is a claim about belief *formation*. Gilbert does not, of course, doubt that subsequent evaluation may lead one to tag a belief as 'false' or declare the status of a claim to be 'unknown'; on his account, it is entirely possible that subsequent (post initial encoding) evaluation could tag a claim as 'plausible'. Experimental attempts to investigate Gilbert's account must consequently target belief formation itself. In order to isolate belief formation, Gilbert employed a dual-task design: if a secondary task disrupts participants' evaluation of novel claims presented to them, then the initial encoding should be all there is, and if that initial encoding consistently renders claims 'true', then Gilbert's account is confirmed.

This is the logic underlying Gilbert et al.'s (1990) "Hopi Language Experiment" on which we based our current studies. In the learning phase of this experiment, participants were presented with a series of statements about the meanings of 'Hopi words.'[1] Importantly, all of these statements were entirely novel, so that participants could not have encoded (and evaluated) them before. Sometimes, a tag indicating whether the preceding sentence was

[1] Hopi is a North American Indian language of the Uto-Aztecan family, spoken by the Hopi people in northeastern Arizona. However, Gilbert et al. used nonsense words under the guise that they were Hopi.

'true' or 'false' followed its presentation. On some occasions during this learning phase, participants were additionally required to perform a secondary task (pressing a key when hearing a tone). Then, a test phase followed the learning phase. Here participants were shown a series of statements, comprising the ones for which truth-values had been provided during the first phase, and additional new statements not seen during the first phase (foils). For each statement, participants had to choose between four response options: 'true', 'false', 'never seen', or 'no information'. 'Never seen' meant that the statement had not been presented during the learning phase. 'No information' meant that the statement had been shown, but there was no indication of its truth-value (since there were no such statements in the test phase, answering 'no information' was always an error).

As predicted, Gilbert et al. (1990) found that judgements about statements that participants had learnt under high cognitive load (i.e., when performing a disrupting secondary task) showed a 'truth bias'. Indeed, disruption did not affect the identification of true statements, but participants misclassified more false statements as true than the reverse when they had learnt them under cognitive load. According to Gilbert et al., cognitive load interrupts the processing of the proposition, which results in participants' bypassing the assessment stage (reading the true/false signal word). Consequently, participants store the proposition as per the default state, that is, 'true'. Hence, interruption of the processing of a claim caused people to mistake false claims for true ones, but not vice versa.

In order to control participants' assessment of the veracity of the propositions, Gilbert et al. (1990) intentionally used "nonsense propositions" of the form *An X is a Y*, where *X* is supposed to be a Hopi noun (but is a nonsense word), and *Y* is supposed to be its English equivalent (e.g., *A twyrin is a doctor*). These propositions do not bear upon the real world and cannot be assigned any prior plausibility by design. As such, they are not representative of most of the propositions we encounter in real life.

In this artificial context, one could explain Gilbert et al.'s results by hypothesising that listeners tend to trust someone's assertion in the absence of any specific reason to doubt.[2]

[2] This view is also advocated by Street and Kingstone (2017), who further emphasise the role of contextual information about the source. In the absence thereof, they claim, the per default assumption is that people are rather trustworthy, which results in a truth bias *when* participants are forced to choose. However, their manipulation does not consist in varying such contextual information but in offering participants the choice to answer '*I don't know*', which cancels the effect. They conclude that participants' answers are their *best guess* rather than reflecting the belief acquired during the first phase.

Human beings would not sustain a communication system (let alone one as complex as language) were it not true that, by and large, other people were likely reliable communicators (see e.g., Hahn, Merdes, & von Sydow, 2018).[3] The claim that people assume that a novel statement from an unknown source is more likely true than not, however, is by no means equivalent to Gilbert et al.'s claim that acceptance is a necessary component of comprehension for any given proposition.

There have been other critiques of Gilbert et al.'s (1990) evidence. Sperber et al. (2010) highlight that the propositions in Gilbert et al.'s study are irrelevant to the participants, who have little reason to be vigilant about them. Relatedly, Hasson et al. (2005) show that automatic acceptance does not obtain when the statement's false version is informative. Richter et al.'s (2009) criticism also relies on the observation that the propositions participants have to learn concern "pseudo-facts" that are "not related to any knowledge or beliefs that they could hold" (p. 539). They further show that the truth bias no longer operates whenever the statements are inconsistent with the participants' background beliefs. Drawing from those studies, Mercier (2017) argues for the existence of a "plausibility checking" mechanism, which would detect inconsistencies between background beliefs and communicated information. According to Mercier, such a mechanism would normally operate when the information is being understood and does not require prior acceptance. But in Gilbert et al.'s (1990) study, nothing in the communicated information allows for such a plausibility check.

In line with those criticisms, the first question we aimed to test in the current research was whether Gilbert et al.'s (1990) 'truth bias' effect would obtain with propositions involving real-world materials. Moreover, we investigated the role that plausibility plays in belief formation. Thus, we used statements bearing on the real world, to which participants could assign some degree of plausibility based on their background knowledge and beliefs (unlike Gilbert et al.'s materials). However, in contrast with Richter et al.'s (2009) materials (e.g., 'Soft soap is edible'), we selected *entirely novel* statements, of which participants were unlikely to have

[3] The claim that people are rather trusting than not when there is no specific reason to doubt either the content of the message or the reliability of the source is widely held and has been one of the central issues in the epistemology of testimony (Coady, 1973; Lackey and Sosa, 2006; Gelfert, 2014). There exists a wealth of literature in cognitive, developmental, as well as social psychology on epistemic vigilance and trust (Clément et al. 2004; Mascaro & Sperber 2009; Sperber et al. 2010; Harris 2012; Harris & Lane 2013; Mills 2013; Harris et al. 2018), as well as the evaluation of others' reliability (Fiske 2018), and the role of the perceived credibility of sources in the evaluation of argument strength and subsequent revision of belief (Hahn et al. 2009).

prior knowledge (like Gilbert et al.'s materials). To this end, we pretested the statements for both relative plausibility and prior knowledge. That ensured that participants would not already have 'tagged' these propositions as true or false (which might explain why default acceptance does not show in Richter et al.'s 2009 study). Hence, we apply 'plausibility' to entirely novel statements. Plausibility is thus a measure of our statements' *a priori* 'believability,' (given background knowledge) rather than the degree to which a previously encountered statement is already believed to be true or false.[4] This is crucial as we want to explore the role of plausibility in belief *formation*.

In an initial, exploratory study, we manipulated cognitive load (half of the learning trials were disrupted) to replicate Gilbert et al.'s (1990) setup as closely as possible. We present the results of this pilot study in 'Supplementary Materials 1' (SM1). We then decided to test our refined explanation for the pattern of results observed in this pilot study (SM1) by running a high-powered, pre-registered confirmatory study (Study 1, *https://osf.io/mkrvd*). We further replicated the results of Study 1 in another pre-registered study (Study 2, *https://osf.io/538zj*). Studies 1 and 2 are the focus of the current manuscript. The critical conditions for which our predictions differ from Gilbert et al.'s (1990) account are the 'disruption' ones. Hence, we chose to drop the 'no disruption' conditions in these two studies (all trials were under cognitive load).

The central claim investigated in the current paper is that the relative plausibility of novel propositions plays a role in encoding and subsequent classification as true or false. Specifically, we hypothesise that the plausibility of novel propositions about the real world impacts people's beliefs towards them. All other things being equal, people should be less prone to believing implausible statements than plausible ones. Consequently, where cognitive load disrupts encoding, people's answers should reflect relative plausibility. Thus, there should be a greater number of correct classifications when truth-value is in line with plausibility (i.e., a greater number of correct answers for plausible true and implausible false statements than for implausible true and plausible false statements). Therefore, we predicted

[4] Our use of 'plausibility' also contrasts with Fazio et al.'s (2019), whose study aims at testing the role of relative plausibility in the so-called 'illusory truth effect', whereby repetition increases the likelihood that a statement will be judged as true. They also pretested their materials for prior plausibility, but the statements they use are not novel to the participants, who already have beliefs regarding them. For instance, 'The Earth is a perfect square' is highly implausible and/because it is also known to be false.

that participants would be more likely to misclassify implausible true statements as 'false' than implausible false statements as 'true'. This prediction contrasts with Gilbert et al.'s (1990) theory of default acceptance. Indeed, in their view, implausible false statements should be remembered as 'true' more frequently than implausible true statements are remembered as 'false'. However, the precise predictions of our studies are complicated by the fact that some statements might be more memorable than others.

Indeed, the implausible true statements in our studies most often refer to 'strange' facts or laws (see Study 1, Materials). Consequently, we reasoned that the *surprise* caused by the learning signal (learning that such an implausible statement is true) might attract greater attention. This would result in participants' tendency to correctly remember those statements as true despite the effect of (im)plausibility posited above. On the other hand, we did not expect such a surprise effect to arise for plausible false statements. Indeed, the scores of our plausibility pretest were less extreme for plausible statements than for implausible ones (participants rated the plausible statements less plausible than they rated the implausible ones implausible; see Study 1, Materials).[5] Consequently, learning that a plausible statement was false was, on average, less surprising than learning that an implausible statement was true.

Despite the complexities outlined above, precise predictions can be made — facilitated by post hoc analysis of our pilot study (SM1). Notably, Gilbert et al. (1990) measured participants' performance for statements for which veracity information was provided ('critical statements', see Design) in terms of two different scores: *correct identification* (of true/false statements as true/false), and *reversals* (of true/false statements as false/true). Although Gilbert et al. (1990) reported complementary patterns of results for correct identifications and reversals, the two need not be complementary. Indeed, participants have other answer options than 'true' and 'false' (namely 'no information' and 'never seen'). Hence, whereas correct identification is the unique measure of success, failure can take different forms, and reversals are one type of mistake among several possible. Furthermore, the results from our

[5] This asymmetry in the plausibility levels of our statements did not appear as a problem for our setup: plausible statements, being closer to 'as likely as not' statements (than implausible ones), are the closest equivalent, among statements bearing upon the real world, to Gilbert et al.'s (1990) statements, which in fact allows us to compare the pattern of our results for plausible statements with Gilbert et al.'s more straightforwardly. See Materials.

pilot study (SM1) aided our understanding that the surprise caused by the signal for implausible true statements might not affect correct identifications and reversals symmetrically. Indeed, the surprise of learning that an implausible statement is true might make participants remember it better (improving their correct identification score). But there is no reason why such surprise should specifically reduce the proportion of reversals (as distinct from other types of errors). Thus, reversals are the key data in this study. For reversals, we thus expected an interaction between veracity and plausibility. Specifically, the proportion of reversals would be greatest for statements for which plausibility was in opposition with truth-value (plausible false and implausible true). Our specific, pre-registered (*https://osf.io/mkrvd*) predictions (ultimately borne out in our data) were:

> [P1] An effect of veracity on plausible statements for reversals (we expected participants to misclassify more plausible false statements as true than plausible true as false);
>
> [P2] An effect of veracity on implausible statements for reversals (we expected participants to misclassify more implausible true statements as false than implausible false as true);
>
> [P3] An effect of plausibility on false statements for reversals (we expected participants to misclassify more plausible false than implausible false statements as true);
>
> [P4] No significant effect of plausibility on true statements for reversals (since implausible true statements are more memorable, there will be few wrong answers).

For plausible statements, our predicted pattern for reversals is identical to Gilbert et al.'s (1990). For implausible statements, our prediction is opposite to Gilbert et al.'s and cannot be explained on their account — assuming that participants have engaged with the learning phase. However, our predictions for reversals are indistinguishable from what would happen if participants were just guessing implausible statements to be false and plausible statements to be true. Hence, we need to make sure that participants actually learned something during the learning phase. Correct identifications scores, specifically participants' good performance with implausible true statements, can provide evidence for this. Indeed, if participants didn't learn anything and only answered in line with plausibility, they would answer correctly for all plausible true and implausible false statements and incorrectly for all implausible true and

plausible false ones (in the absence of response noise). Thus, any correct recall of implausible true and plausible false statements suggests some learning has taken place. Specifically, invoking the notion of 'surprisingness' highlighted earlier, the following predicted results will demonstrate that learning has indeed occurred:

> [P5] An effect of veracity on plausible statements (we expected more correct identifications of plausible true statements than plausible false);
>
> [P6] No effect of veracity on implausible statements (we expected no difference between correct identifications of implausible true and implausible false statements, because of the 'surprise' effect);
>
> [P7] An effect of plausibility on false statements (we expected more correct identifications of implausible false than plausible false statements);
>
> [P8] No effect of plausibility on true statements (we expected no difference between correct identifications of implausible true and plausible true statements because of the surprise effect for implausible statements).

Hence, although our key data are reversals, results for correct identifications are indispensable in demonstrating that we are testing the effect of plausibility on *encoding* propositions rather than its effect on *evaluating* them when reading them for the first time. In brief, the surprise caused by the learning signal may enhance participants' memorisation of statements and their truth-value. But, when participants do not correctly remember statements, their answer should reflect prior plausibility rather than default acceptance.

Study 1

Method

Participants

106 participants (48 female, 58 male) were recruited and paid via Amazon Mechanical Turk (M_{age} = 39 [21-70]; *SD* = 11,16). They were paid $4,50 for this ~30-minute task. The study

received approval from the Dept. of Psychological Sciences Ethics committee at Birkbeck College.

Design

Twenty-four *critical statements* were divided into a 2x2 repeated measures design. The independent variables were plausibility (Plausible [P] or Implausible [I]) and veracity (True [T] or False [F]). For counterbalancing, each critical statement occurred in a 'true' version and a closely related 'false' version. We thus generated two versions of the questionnaire with opposing veracities for the critical statements, such that each participant only saw one version of each critical statement. In addition to these 24 critical statements, participants saw 12 'fillers' and 12 'buffers' (either true or false, but which did not come in pairs — all participants saw the same); hence each participant saw 48 statements in the 'learning phase'. 'Fillers' were statements about which no truth-value information was provided (followed by a blank screen); they did not appear during the test phase. We used fillers to replicate Gilbert et al.'s (1990) setup as closely as possible. Specifically, Gilbert et al. included fillers to make it more difficult for participants to "keep track of how many instances of each signal word they had seen and thus inhibited the tendency to respond *true* and *false* to approximately equal numbers of test items" (p. 603). Fillers also helped to increase the credibility of Gilbert et al.'s cover story. 'Buffers' were statements either followed by truth-value information or not, shown at the very beginning and the very end of the learning phase. Buffers were intended to avoid primacy and recency effects and did not appear during the test phase.

Our dependent variables were the scores for correct identifications and reversals. *Correct identification* means classification of a true statement as true or a false statement as false. *Reversal* is a specific type of error, referring to the misclassification of a true statement as false or a false statement as true. Both concern critical statements (for which there was an indication of truth value) only. We calculated scores from participants' answers to critical statements in the test phase, where they had to choose between four response options: 'true', 'false', 'never seen', and 'no information'. *No information* was the appropriate response for the statements about which no truth-value information was provided during the learning phase (those followed by a blank screen, called 'fillers'). Since none of them appeared during the test phase, 'no information' was always a wrong answer. *Never seen* was the

appropriate response for statements that were entirely novel in the test phase (foils). 'Never seen' was always a wrong answer for critical statements. Since we were primarily interested in correct identifications and reversals (like Gilbert et al. 1990), we focused on critical statements only and did not distinguish, in our analyses, between 'no information' and 'never seen' responses.[6]

Materials

Our materials consisted of two sets of 48 statements, which had to be novel for participants, as in Gilbert et al.'s (1990) 'Hopi language' experiment (to avoid situations where participants have those propositions already 'tagged' as true or false), but with some level of plausibility.

To this end, we chose statements about the real world, primarily laws of foreign countries, as well as demographical, historical, geographical and zoological facts. We selected them after pretesting for plausibility and prior knowledge. Note that there was no deception in this study: to the best of our knowledge, the statements indicated as true were true, and those indicated as false were false. See 'Supplementary Materials 2' for the full list of statements.

Pretesting

We pretested 160 statements (80 pairs of statements). Each pair consisted of one true statement and one closely related false statement. From a web inquiry, we initially drew a list of supposedly (as far as a quick check on the Internet confirmed[7]) true statements that appeared novel (we didn't expect most people to have prior knowledge of them). Some of them sounded implausible ('strange facts', *e.g. Female kangaroos have three vaginas*, or 'strange laws', *e.g. In Florida, the law forbids single women to parachute on Sundays*), while others sounded plausible (*e.g. In Denmark, drivers are supposed to drive their vehicles with their headlights on during the day as well*). From this initial list of 80 true statements (of varying levels of plausibility), we generated a list of 80 false statements, intended to be equivalent in terms of plausibility (*e.g.* by changing the name of the place or animal concerned: *Female wallabies have three vaginas*; *In South Carolina, the law forbids single*

[6] As an anonymous referee highlighted, there was a fifth option (and a fourth way of being wrong), namely failing to choose any of the proposed four answers. We did not distinguish this kind of (non)-answers from 'no information' and 'never seen' answers. Such answers did not occur in Study 1 and were rare in Study 2 (1.5% of the critical trials, more than half of those -60% - from the same participant).

[7] E.g., http://edition.cnn.com/2009/TRAVEL/04/03/worlds.strangest.laws/?iref=nextin; https://www.nytimes.com/2007/08/07/world/asia/07cnd-thai.html

women to parachute on Sundays; In Norway, drivers are supposed to drive their vehicles with their headlights on during the day as well).

Twenty participants (recruited via Amazon Mechanical Turk) each rated one statement of each pair (either true or false) for plausibility on a 100-point scale ("How likely do you think this statement is to be true?"). They subsequently indicated whether they had already known this.

We excluded all statements for which at least one participant indicated knowledge (except where this indication was obviously false — because the participant rated a false statement as very likely, or a true statement very unlikely). Among the statements that were unknown to all participants, we selected 24 pairs as our *critical statements*. These included six (pairs of) statements in each of the following four categories: 'plausible true', 'plausible false', 'implausible true', 'implausible false'). The two versions (one false, one true) of all selected statements had received a plausibility rating below 27 or above 56 (following a visual inspection of the distribution of plausibility ratings).

Hence, as mentioned in the Introduction, there was an asymmetry in the plausibility levels of our statements: implausible statements were rated much more implausible (below 27) than plausible statements were rated plausible (above 56). Indeed, it turned out to be practically impossible to come up with extremely plausible statements, which would nevertheless be novel in the same sense as our implausible statements were. For example, consider the negation of one of our implausible statements, "In Ireland, it is allowed for women to eat chocolate on public transport." There is a sense in which, even though one has never thought about this claim, one already *knows* it (which is why its negation is surprising). In other words, a highly plausible statement seems to be less informative than its negation, even though both concern facts about which one has never thought before. We will not further elaborate on this subtle issue related to fundamental problems about informativeness and negation (see Hasson et al. 2005). Suffice it to say that the asymmetry in plausibility levels was not a problem for our setup. Quite the contrary, since our 'plausible' statements were the closest equivalent, among statements bearing upon the world, to Gilbert et al.'s a-plausible statements.

In addition to the 24 pairs of critical statements, we further selected 36 statements to be used as fillers, buffers, and foils.

Procedure

The procedure was adapted from Gilbert et al.'s (1990) Study 1 ("The Hopi Language experiment"). The main difference to their procedure concerned disruption: we used a different secondary task[8], and all trials were disrupted. As explained in the Introduction, the disrupted conditions are where we expect to see the effects of our independent variables, Plausibility and Veracity. They are where our predictions differ from Gilbert et al.'s predictions. Therefore, we conducted all trials under cognitive load to maximise the reliability of our findings in the critical 'disruption' conditions. In addition, the study was conducted online instead of in the lab.

Participants were invited to take part in an experiment about how people learn facts. Similar to Gilbert et al.'s (1990) participants, our participants were told that "the experiment is designed to simulate the flow of novel information that people are confronted with in their everyday life. In some cases, information gets confirmed later on; in others, it gets falsified. Our goal is to study how people deal with that."

The experiment was divided into two phases, as in Gilbert et al.'s (1990) Hopi Language study.

Learning phase

After providing informed consent, participants were told that they would read a series of statements, some of which would be followed by a tag indicating their truth-value (TRUE or FALSE). It was made clear that there would be no deception about truth-value (signal words TRUE / FALSE always follow true/false statements). Participants were informed that the trial statements that were not followed by any truth-value indication helped simulate a more real-life flow of information. Participants were further told that they would have to perform an additional, secondary task (typing a sequence of numbers they would have just heard — our disruption task). After short practice of the disruption task, participants saw the 48 statements on the computer screen, presented in a random order[9], one at a time.

Each statement appeared for 8 seconds, followed by a 2-second blank screen, followed by a

[8] As explained in 'Supplementary Materials 1', the distraction task used by Gilbert et al. (1990) was not enough to distract our participants (it had no effect at all), and we had to run several pilots until we found a disruption task that impaired performance.

[9] The one constraint was that the first and last six statements were buffers.

voice from the computer giving a random sequence of 5 digits (e.g., 6-0-7-4-5). Immediately after hearing the sequence, participants saw either a blank screen or a signal word TRUE or FALSE for 1 second. Then a sentence on the computer screen appeared, asking the participant to type the sequence of digits they had just heard. Participants were allowed 8 seconds to type the sequence. Then the next statement appeared.

Test phase

During the test phase, participants saw the 24 critical statements (6 plausible true, 6 plausible false, 6 implausible true, 6 implausible false) and 12 foils (propositions not shown during the learning phase). Statements appeared one at a time. 'Fillers' were not presented during the test phase. As in Gilbert et al.'s study, participants were allowed 9 seconds to answer by clicking on any of the following four buttons: *true*, *false*, *no information*, and *never seen.*

Statistical analyses

The software R (version 4.1.1; R Core Team, 2021) was used for all analyses. In particular, the packages *afex* (Singmann, Bolker, Westfall, & Aust, 2016) and *lme4* (Bates, Mächler, Bolker, & Walker, 2015) were used for ANOVAs and mixed-effects logistic regressions, respectively.

Results

Planned analyses

A visual inspection of the data is in line with the predicted results for both reversals (Figure 1) and correct identifications (Figure 2). Proportions of reversals were greater for plausible false than plausible true statements, but participants reversed more implausible true than implausible false statements. In addition, participants reversed more plausible false than implausible false statements and, finally, more implausible true than plausible true (not predicted). Proportions of correct identifications were greater for plausible true than for plausible false (following Gilbert et al.'s pattern), as well as for implausible false than for

plausible false, but they were equivalent for implausible false, implausible true, and plausible true.

The proportion of false alarms (i.e., when participants responded with 'False' or 'True' to a foil) was 24.1%. 90% of answers to the distractor task were correct.

Since reversals are our key data, we present the inferential statistics for reversals first.

Number of reversals

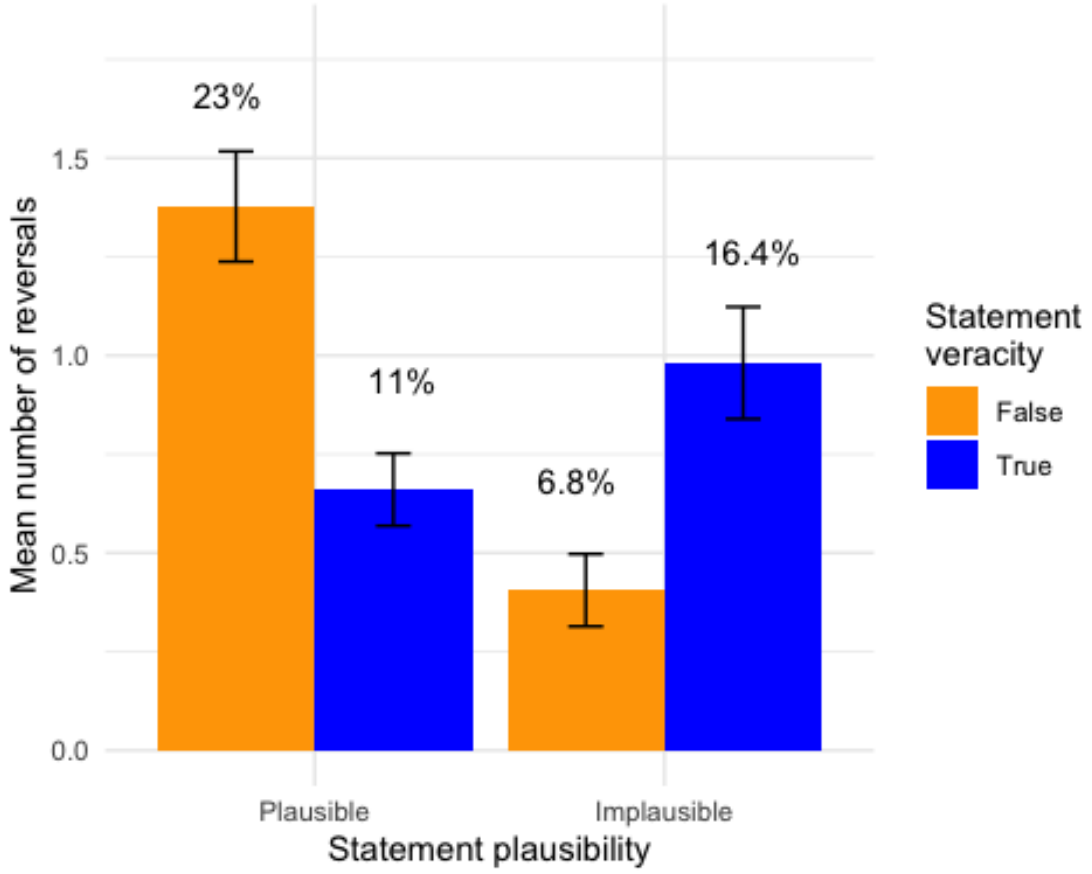

Fig. 1 Reversals (absolute number shown on the x-axis and written as a percentage of trials). Error bars are plus and minus one standard error.

A 2 (veracity) x 2 (plausibility) repeated measures ANOVA revealed the expected statistical interaction of plausibility and veracity, $F(1,105) = 27.75$, $p < .001$, $\eta_p^2 = .209$, with a main effect

of plausibility, $F(1,105) = 19.81$, $p < .001$, $\eta_p^2 = .159$, but no main effect of veracity, $F(1,105) = 0.60$, $p = .439$, $\eta_p^2 = .006$.

One-way ANOVAs were subsequently conducted to test the specific predictions [P1-P4]. They confirmed that: there were significantly more reversals for false than true plausible statements [P1], $F(1,105) = 22.93$, $p < .001$, $\eta_p^2 = .179$, and significantly more reversals for true than false implausible statements [P2], $F(1,105) = 13.66$, $p < .001$, $\eta_p^2 = .115$. As predicted, there were significantly more reversals for plausible than implausible false statements [P3], $F(1,105) = 56.22$, $p < .001$, $\eta_p^2 = .179$. Finally, there were also significantly more reversals for implausible than plausible true statements [not predicted, P4], although this effect was weaker than the other effects, $F(1,105) = 4.29$, $p = .041$, $\eta_p^2 = .039$.

In addition to these pre-registered analyses, which enable a ready comparison with Gilbert et al.'s (1990) original study, we verified these results with an unplanned mixed-effects logistic regression on the likelihood of reversal (compared with any other answer category), predicted by veracity, plausibility and their interaction. Such an analysis respects the categorical nature of responses for each statement.[10] Our model allowed for random effects (intercept and slope) based on repeated measures from subjects and variations within item presentation as either true or false[11]. The results were consistent with those from the ANOVAs. Again, we found a significant interaction between veracity and plausibility, $b = -0.80$, $SE = 0.20$, $z = -3.94$, $p < .001$, a significant main effect of plausibility, $b = 0.86$, $SE = 0.21$, $z = 4.05$, $p < .001$, but no main effect of veracity, $b = 0.29$, $SE = 0.20$, $z = 1.42$, $p = .155$. Unpacking the interaction[12], for plausible items, true statements were less likely to be reversed than false statements [P1], b = -0.58, SE = 0.17, z = -3.53, p < .001; for implausible items, true statements were more often reversed than false statements [P2], b = 1.30, SE = 0.56, z = 2.32, p = .020; for false statements, plausible items were more often reversed than implausible items [P3], $b = 1.65$, $SE = 0.34$, $z = 4.8$, $p < .001$. The only result which differed from the original ANOVA, was that the

[10] We thank an anonymous reviewer for suggesting this analysis. Associated figures (which show the same pattern of results as Figure 1) can be found in Supplementary Materials 3.

[11] The true statement and its generated false statement, intended to be equivalent in terms of plausibility (e.g., "*In Florida, the law forbids single women to parachute on Sundays"; "In South Carolina, the law forbids single women to parachute on Sundays")* were coded as being variations of the same item. The exact model used was: glmer(reversal ~ veracity * plausibility + (veracity * plausibility|subject) + (veracity|item)).

[12] The exact models tested were glmer(reversal ~ veracity + (veracity|subject) + (veracity|item)) and glmer(reversal ~ plausibility + (plausibility|subject)).

unpredicted effect of plausibility for true statements was not observed in this analysis, b = 0.07, SE = 0.18, z = 0.42, p = .677 [consistent with P4].

Number of correct identifications

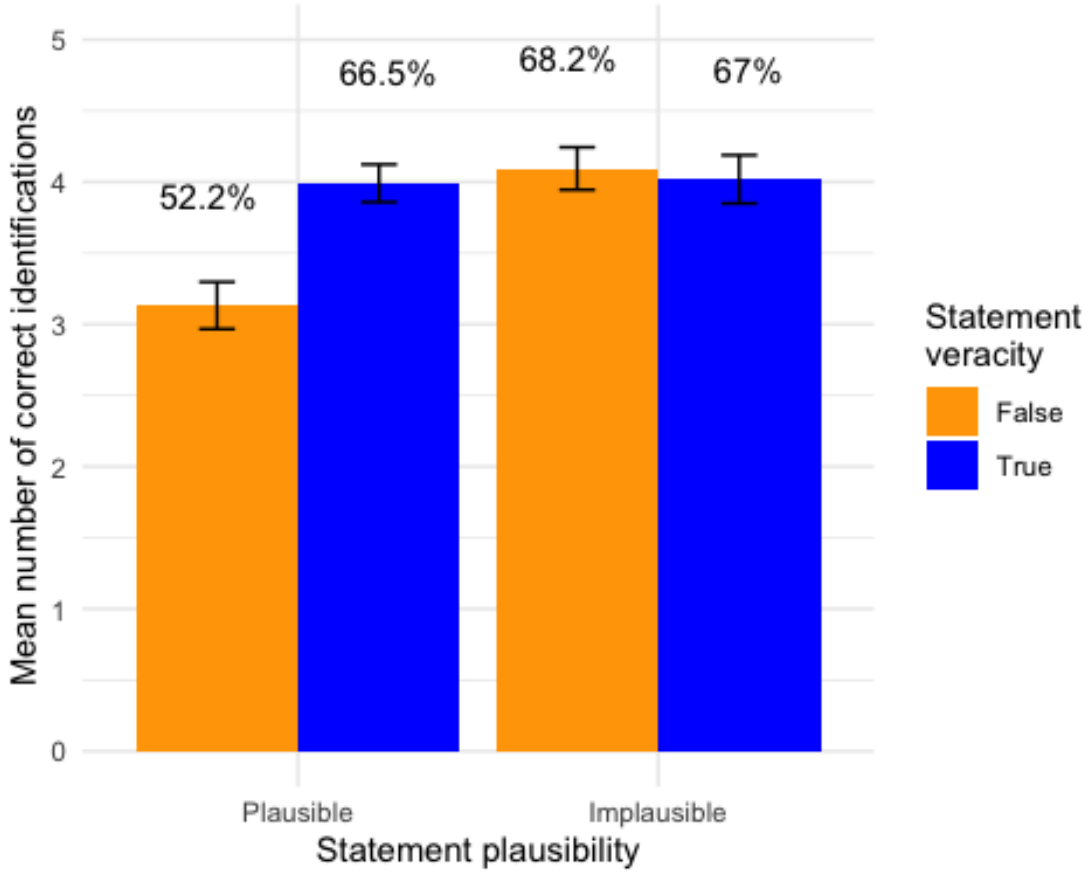

Fig. 2 Correct identifications (absolute number shown on the x-axis and written as a percentage of trials). Error bars are plus and minus one standard error.

A 2 (veracity) x 2 (plausibility) repeated measures ANOVA revealed a statistical interaction of plausibility and veracity for correct identification, $F(1,105) = 9.29$, $p = .003$, $\eta_p^2 = .081$. We found main effects of plausibility, $F(1,105) = 15.76$, $p < .001$, $\eta_p^2 = .131$, and of veracity, $F(1,105) = 10.53$, $p = .002$, $\eta_p^2 = .091$.

One-way ANOVAs were conducted to test the specific predictions [P5-P8]. They confirmed that there were more correct identifications for true than false plausible statements [P5], $F(1,105) = 20.77$, $p < .001$, $\eta_p^2 = .165$, but no difference between true and false implausible statements [P6], $F(1,105) = 0.14$, $p = .709$, $\eta_p^2 = .001$. As predicted, there were also more correct identifications for implausible than plausible false statements [P7], $F(1,105) = 23.98$, $p < .001$, $\eta_p^2 = .186$, but no difference between implausible and plausible true statements [P8], $F(1,105) = 0.02$, $p = .887$, $\eta_p^2 = .000$.

We again ran an (unplanned) mixed-effects logistic regression to verify these results.[13] There was a significant interaction between plausibility and veracity, $b = 0.16$, $SE = 0.08$, $z = -2.07$ $p = .038$, and significant effects of plausibility, $b = -0.27$, $SE = 0.10$, $z = 2.76$, $p = .006$, and veracity, $b = 0.16$, $SE = 0.07$, $z = -2.39$, $p = .017$. Following the same approach as for reversals, follow-up analyses showed that: for plausible items, true statements were more likely to be correct [P5], $b = 0.32$, $SE = 0.09$, $z = 3.5$, $p < .001$; for implausible items, there was no difference in the likelihood for correct identification of true versus false statements [P6], $b = -0.01$, $SE = 0.11$, $z = -0.10$, $p = .918$; for false statements, implausible items were more likely to be correct than plausible items [P7], $b = -0.40$, $SE = 0.08$, $z = -4.83$, $p < .001$; for true statements, there was no difference in the likelihood of correctness for plausible versus implausible items [P8], $b = -0.11$, $SE = 0.09$, $z = -1.15$, $p = .252$.

Exploratory analyses

Further to our planned analyses, we ran exploratory analyses, intended to offer another way of looking at our results and further confirming the conclusions from the preceding analyses. This new way of considering our results centres on the notion of *familiarity*.

Participants who respond with 'True' or 'False' to learnt (critical) statements (as opposed to foils), independently of whether this was a correct identification or a reversal, suggest familiarity with these statements. Indeed, a participant's responding either 'True' or 'False' means that they (correctly) remember having been presented with the statement and taught

[13] glmer(correct identification ~ veracity * plausibility + (veracity * plausibility|subject) + (veracity|item)).

something about its truth-value. Otherwise, they would select the 'Never seen' or 'No information' answer. Both correct identifications and reversals (which together correspond to all 'True' or 'False' response choices to critical statements) therefore indicate familiarity – that is, remembering the statement as a previously learnt one. Only in the case of reversals do participants *mis*remember the truth-value — they misclassify a true/false statement as false/true. On the other hand, responses of 'No Information' or 'Never Seen' suggest that participants forgot that they had learnt something about these statements altogether — that is, these statements are unfamiliar to them.

As explained above (see 'Design'), a third (33.0%) of statements presented in the test phase had not appeared in the learning phase (foils). Here, however, we focus on critical statements only. Overall, participants responded to 22.2% of critical statements as 'unfamiliar' (or false negatives — 'no information' or 'never seen'). Figure 3 shows the proportions of familiar statements (in blue), with a darker blue for reversals.

Focussing on both blue regions of Figure 3, it appears as though implausible true statements are most likely to be recognised as familiar, in line with our notion of surprisingness: learning that an implausible statement is true might boost its memorability. Focussing solely on the dark blue regions, implausible true statements have more reversals than plausible true and implausible false statements. Indeed, among the high number of implausible true statements that are *familiar*, there is a high proportion of reversals. This is in line with our prediction that surprise should not lessen the proportion of reversals specifically (by contrast with other types of wrong answers). On the other hand, plausible false statements have a lower score in terms of familiarity. And they have the highest score in terms of reversals alone, which means that they have the lowest score by far in terms of correct identifications.

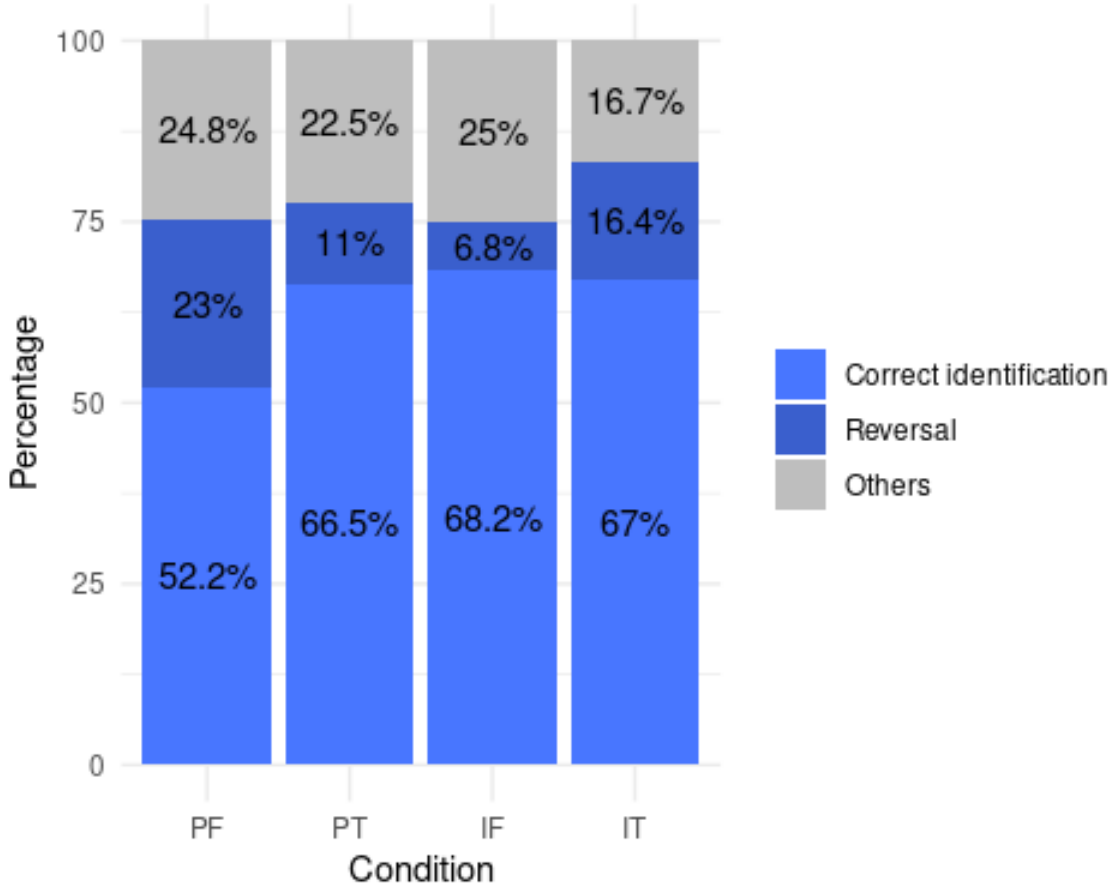

Fig. 3 Proportion of correct identifications and reversals (i.e., 'familiar', in blue) and other answers (i.e., 'unfamiliar', in grey) for critical statements only (excluding foils).

Likelihood of familiar classification

The above observations would be born out statistically with a veracity × plausibility interaction such that implausible true statements were more likely to be recognised than any other statements. A mixed-effects logistic regression on the likelihood of a statement being seen as familiar ('true' or 'false') vs unfamiliar ('never seen', 'no information')[14] did not, however, reveal a significant interaction, b = -0.14, SE = 0.08, z = -1.72, p = .086. There were also no significant effects of veracity, b = 0.12, SE = 0.08, z = 1.47, p = .141, or plausibility, b = -0.17, SE = 0.11, z = -1.46, p = .144.

Likelihood of correct identification vs reversal (within 'familiar' statements)

[14] glmer(familiarity ~ veracity * plausibility + (veracity * plausibility|subject) + (veracity|item)). Figures presenting estimated P(familiarity) and P(correct identification|familiarity), which show the same pattern of results as Figure 3, can be found in Supplementary Materials 3.

A mixed-effects logistic regression on the likelihood of correct identification vs. reversal within 'familiar' statements for the predictors plausibility, veracity and their interaction revealed a significant plausibility × veracity interaction, b = 0.80, SE = 0.22, z = 3.64, p < .001 and a significant main effect of plausibility, b = -0.95, SE = 0.23, z = -4.20, p < .001, but not of veracity, b = -0.22, SE = 0.22, z = -1.02, p = .308.

Taken together, the exploratory analyses mirror the planned analyses and show some indication that veracity and plausibility influence whether a statement is remembered (either correctly or not). Indeed, implausible true statements are marginally (though not significantly) more likely to be seen as familiar, presumably because they are surprising. They are then also remembered with a higher probability of correctness.

Study 2 (replication study)

Method

Our replication study, which was also pre-registered (*https://osf.io/538zj*), followed the exact same design and procedure as Study 1; only the materials were partially modified, and the 'exploratory' analyses from Study 1 were pre-planned. We predicted the same pattern of results as described in Figure 3. Specifically, our predictions regarding reversals and correct identification were the same as in Study 1 [P1-8]. In line with the notion of surprisingness, whereby learning that an implausible statement is true might boost its memorability, we expected a veracity x plausibility interaction for likelihood of familiar classification such that implausible true statements were most likely to be recognised as familiar.[15] Last, we expected a significant interaction of veracity and plausibility for likelihood of correct identification vs reversals (within 'familiar' statements), with participants being most likely to make mistakes (reversals) for plausible false and for implausible true statements.

[15] Our initial conviction in this prediction was originally further supported by the fact that we initially observed a significant veracity × plausibility interaction in Study 1, using the model, glmer(familiarity ~ veracity * plausibility + (1|subject)), before we recognised the maximal model subsequent to running Study 2, glmer(familiarity ~ veracity * plausibility + (veracity * plausibility|subject) + (veracity|item)). In Study 2, the two model specifications result in identical patterns of significance, unless highlighted. The one reported here was consequently not the precise specification included in the pre-registration.

Participants

115 participants[16] (67 female, 48 male) were recruited and paid via Prolific Academic (M_{age} = 32 [18-73]; *SD* = 11,55). They were paid $3,13 for this ~25-minute task.[17]

Materials

The materials consisted of the same number of statements. Some were similar to Study 1, while others were introduced after a new pretest. Thirteen pairs of critical statements (of 24) were changed. In particular, we removed statements that had become obsolete (e.g., legislation on same-sex marriage in Australia has changed since our first study), or for which we had discovered that, despite our best efforts, the veracity information we had provided in Study 1 was wrong (e.g., it turns out that female wallabies have three vaginas too). Furthermore, we introduced two pairs of plausible statements about the (real) meaning of some Hopi and Shoshone[18] words. Ratings in the pretest confirmed that statements for which one has no clue regarding their plausibility tend to be judged as rather plausible (ratings for such statements ranged from 54 to 67).[19] See 'Supplementary Materials 2' for the full list of statements.

Finally, we ensured a balanced number of plausible and implausible statements in the non-critical categories (buffers and fillers) and in the foils for Study 2. That allowed us to test whether there was an effect of plausibility on participants' responses to foils.

Results

A visual inspection of the data is in line with the predicted results for both reversals (Figure 4) and correct identifications (Figure 5). Proportions of reversals were greater for plausible false

[16] Nine participants could not finalise the submission of their participation to Prolific Academic, despite their having completed the study and their data being stored. We decided to pay them and use their data, as there was no reason not to. Hence, the number of participants slightly exceeds our preregistered plans.

[17] The replication study was the same length as Study 1, but we had initially overestimated the completion time.

[18] Shoshone is another North American Indian language of the Uto-Aztecan family.

[19] In line with the assumption of per default trust in others' assertions, we can interpret these moderately high ratings of plausibility as a 'why not if you say so' answer rather than a judgement on the intrinsic plausibility of the statement's content.

than plausible true statements, and participants reversed more implausible true than implausible false statements. Participants again reversed more plausible false than implausible false statements and more implausible true than plausible true. Proportions of correct identifications were greater for plausible true than for plausible false (following Gilbert et al.'s 1990 pattern), as well as for implausible false than for plausible false but they were equivalent for implausible false, implausible true, and plausible true (Figure 5). As shown in Figure 6, implausible true statements had a higher familiarity score but, among familiar statements, implausible true (and plausible false) had a higher number of reversals than plausible true and implausible false. The proportion of false alarms (i.e., when participants responded with 'False' or 'True' to a foil) was lower than in Study 1 (16.2%). Eighty-eight percent of the answers to the distractor task were correct.

Unless otherwise indicated, the analyses reported below were all planned.

Number of reversals

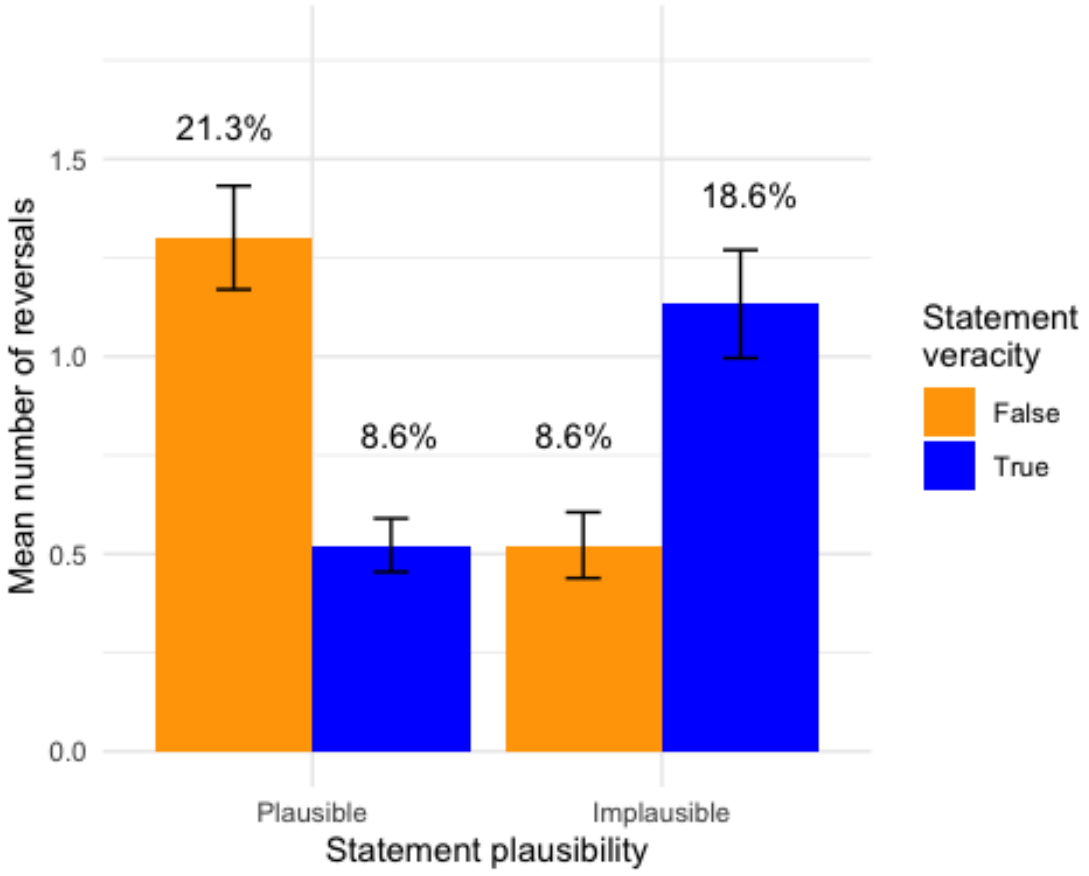

Fig. 4 Reversals (absolute number shown on the x-axis and written as a percentage of trials). Error bars are plus and minus one standard error.

A 2 (veracity) x 2 (plausibility) repeated measures ANOVA revealed the predicted statistical interaction of plausibility and veracity, $F(1,105) = 48.97$, $p < .001$, η_p^2 .304, but no main effect of plausibility, $F(1,105) = 1.2$, $p = .275$, η_p^2 .011, and no main effect of veracity, $F(1,105) = 0.89$, $p = .347$, η_p^2 .008.

One-way ANOVAs were subsequently conducted to test the specific predictions. They confirmed that there were significantly more reversals for false than true plausible statements [P1], $F(1,105) = 32.05$, $p < .001$, $\eta_p^2 = .223$, and significantly more reversals for true than false implausible statements [P2], $F(1,105) = 22.38$, $p < .001$, $\eta_p^2 = .167$. As in Study 1, there were significantly more reversals for plausible than implausible false statements [P3], $F(1,105) = 40.51$, $p < .001$, $\eta_p^2 = .266$, and there were also significantly more reversals for implausible than plausible true statements [not predicted, P4], $F(1,105) = 22.59$, $p < .001$, $\eta_p^2 = .168$.

Again, we tested these results with an (unplanned) mixed-effects logistic regression. Critically, we replicated the interaction between plausibility and veracity, $b = -0.44$, $SE = 0.10$, $z = -4.55$, $p < .001$. We also observed a main effect of plausibility, $b = 0.20$, $SE = 0.10$, $z = 2.04$, $p = .042$, in contrast to the ANOVA based analysis. As in that ANOVA, there was no effect of veracity, $b = -0.03$, $SE = 0.10$, $z = -0.30$, $p = .762$. Unpacking the interaction, for plausible items, true statements were less likely to be reversed than false statements [P1], $b = -0.47$, $SE = 0.14$, $z = -3.33$, $p < .001$; for implausible items, true items were significantly more likely to be reversed than false items [P2], $b = 0.41$, $SE = 0.16$, $z = 2.54$, $p = .011$; for false statements, plausible items were more often reversed than implausible items [P3], $b = 0.63$, $SE = 0.14$, $z = 4.59$, $p < .001$. As in Study 1, the only result which differed from the original ANOVA was that the significant effect of plausibility for true statements was not observed in this analysis [consistent with P4], $b = -0.23$, $SE = 0.13$, $z = -1.77$, $p = .077$.

Number of correct identifications

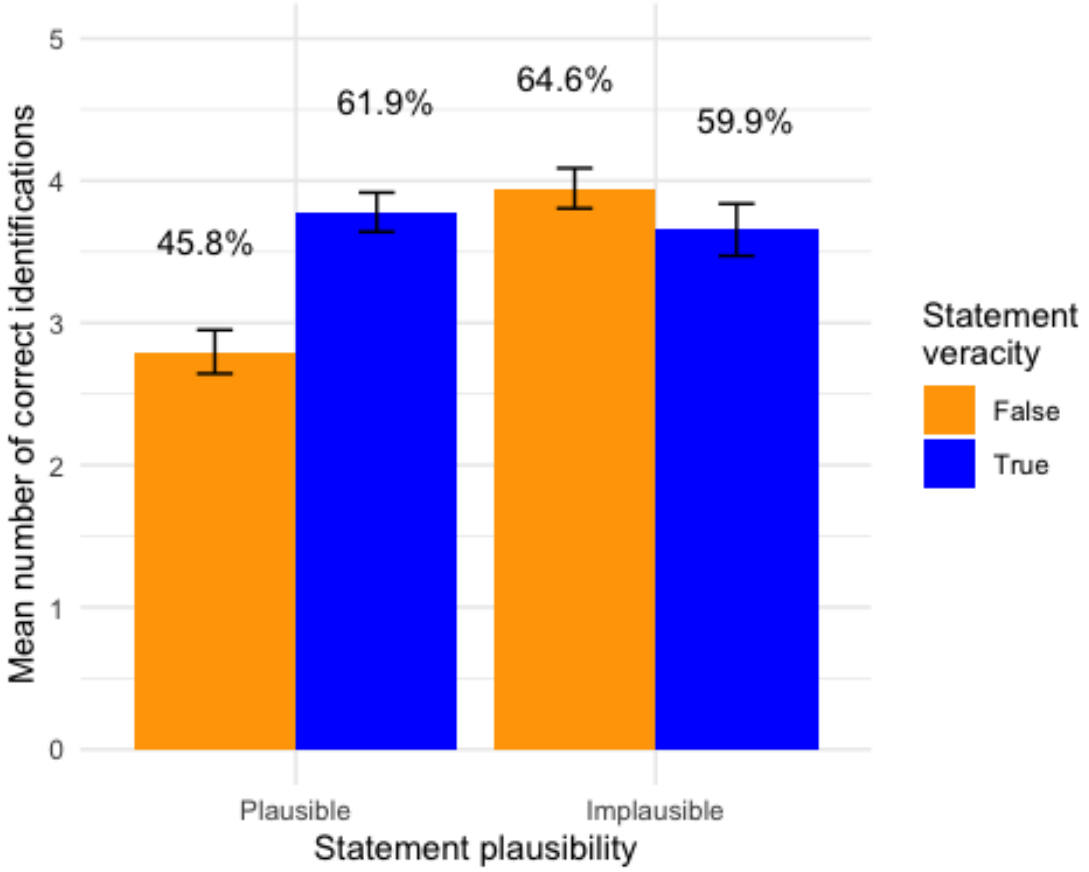

Fig. 5 Correct identifications (absolute number shown on the x-axis and written as a percentage of trials). Error bars are plus and minus one standard error.

A 2 (veracity) x 2 (plausibility) repeated measures ANOVA revealed the predicted statistical interaction of plausibility and veracity for correct identification, $F(1,105) = 25.17$, $p < .001$, η_p^2 .184. As before we found main effects of plausibility, $F(1,105) = 20.81$, $p < .001$, η_p^2 .157, and of veracity, $F(1,105) = 7.79$, $p = .006$, η_p^2 .065.

One-way ANOVAs were conducted to test the specific predictions. They confirmed that there were significantly more correct identifications of true than false plausible statements [P5], $F(1,105) = 30.08$, $p < .001$, $\eta_p^2 = .212$, but no difference between true and false implausible statements [P6], $F(1,105) = 2.77$, $p = .099$, $\eta_p^2 = .024$. As predicted, there were also significantly more correct identifications for implausible than plausible false statements [P7], $F(1,105) = 53.61$, $p < .001$, $\eta_p^2 = .324$, but no difference between implausible and plausible true statements [P8], $F(1,105) = 0.47$, $p = .496$, $\eta_p^2 = .004$.

Again, we tested these results with an (unplanned) mixed-effects logistic regression and confirmed the results from the ANOVAs above. There was a significant interaction between veracity and plausibility, b = 0.23, SE = 0.06, z = 3.58, p < .001, and significant main effects of plausibility, b = -0.24, SE = 0.07, z = -3.39, p < .001, and of veracity, b = 0.15, SE = 0.06, z = 2.37, p = .018. For plausible items, true statements were more likely to be classified correctly than false statements [P5], b = 0.37, SE = 0.07, z = 5.11, p < .001 ; for implausible items, there was no difference in the likelihood for correct identification of true versus false statements [P6], b = -0.09, SE = 0.11, z = -0.76, p = .448 ; for false statements, implausible items were more likely to be correct than plausible items [P7], b = -0.45, SE = 0.06, z = -7.31, p < .001; for true statements, there was no difference in the likelihood of correctness for plausible versus implausible items [P8], b = -0.02, SE = 0.08, z = -0.24, p = .807.

Likelihood of familiar classification

Participants responded to 27.7% of statements as 'unfamiliar' (or false negatives; Figure 6). As predicted, in line with the notion of surprisingness, Figure 6 again suggests that implausible true statements were most often recognised as familiar. As in Study 1, though, the plausibility × veracity interaction was not significant, b = -0.04, SE = 0.06, z = -0.78, p = .433. On this occasion, however, main effects were observed for both plausibility (with implausible statements more likely to be familiar), b = -0.22, SE = 0.08, z = -2.72, p = .007, and veracity (with true statements more likely to be familiar), b = 0.14, SE = 0.06, z = 2.42, p = .015. In summary, participants were more likely to recognise a statement as familiar if it was implausible and if it was true, but these effects were additive rather than multiplicative.

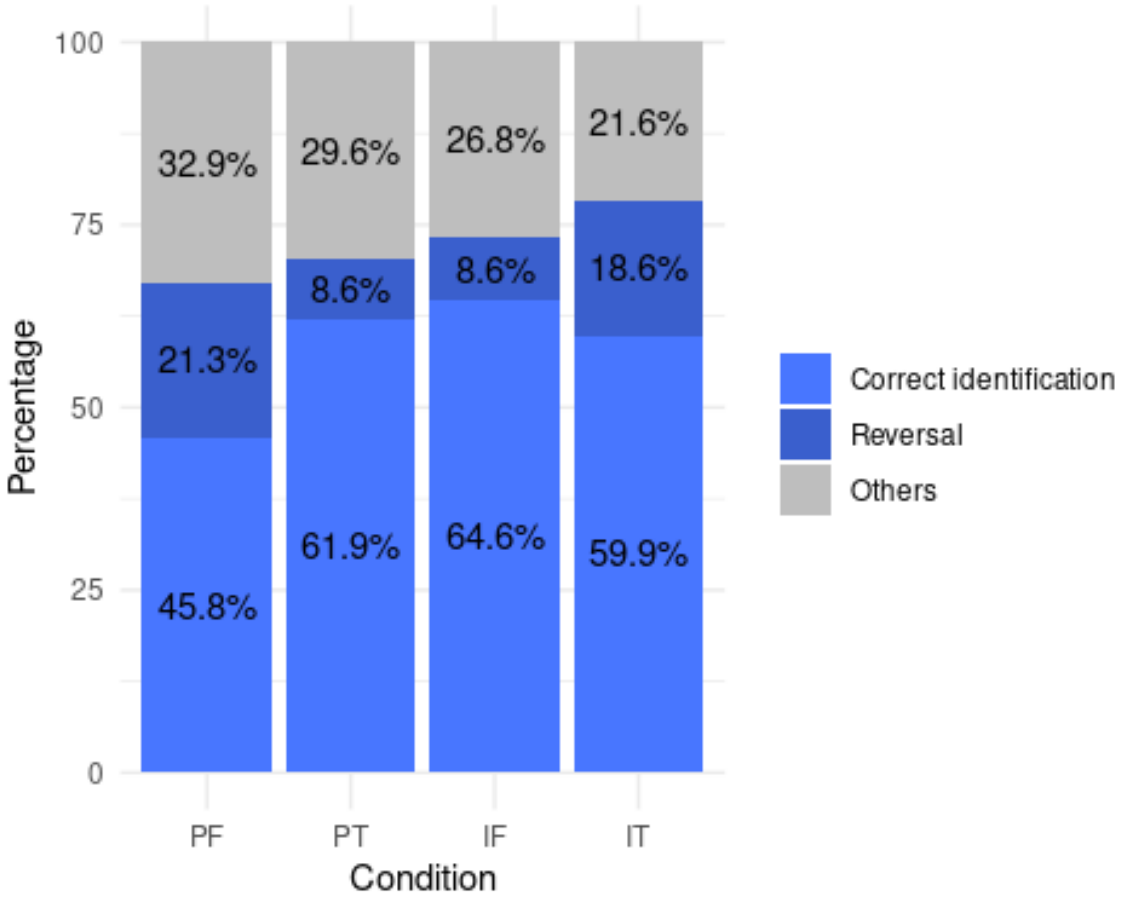

Fig. 6 Proportion of correct identifications and reversals (i.e., 'familiar', in blue) and other answers (i.e., 'unfamiliar', in grey) for critical statements only (excluding foils).

Likelihood of correct identification vs reversal (within 'familiar' statements)

From Figure 6, it is apparent that the highest proportion of reversals was again observed for plausible false statements, but with the next highest proportion observed for implausible true statements (in line with Study 1 and our predictions). As in Study 1, there was a significant plausibility × veracity interaction, b = 0.50, SE = 0.10, z = 4.82, p < .001, and a main effect of plausibility, b = -0.29, SE = 0.11, z = -2.75, p = .006[20]. Contrary to Study 1, there was no significant main effect of veracity, b = 0.10, SE = 0.11, z = 0.89, p = .374

[20] This main effect was not observed in the model only allowing for random intercepts.

Likelihood of correct identification of foils

Finally, we analysed whether there is an effect of plausibility on 'foils', that is, on statements that were not shown during the learning phase, but appeared during the test phase. For all of these, the correct answer is 'Never Seen'. We tested this in a mixed-effects logistic regression with 'Never Seen' (Correct) vs any other answer (Incorrect) as the dependent variable and plausibility of foil statements as the predictor variable. As before, the model allowed for random effects at participant level[21]. The results showed a significant negative effect of plausibility, b = -1.00, SE = 0.16, z = -6.12, p < .001. In other words, participants were more likely to recognise foils as such when they were implausible (Figure 7).

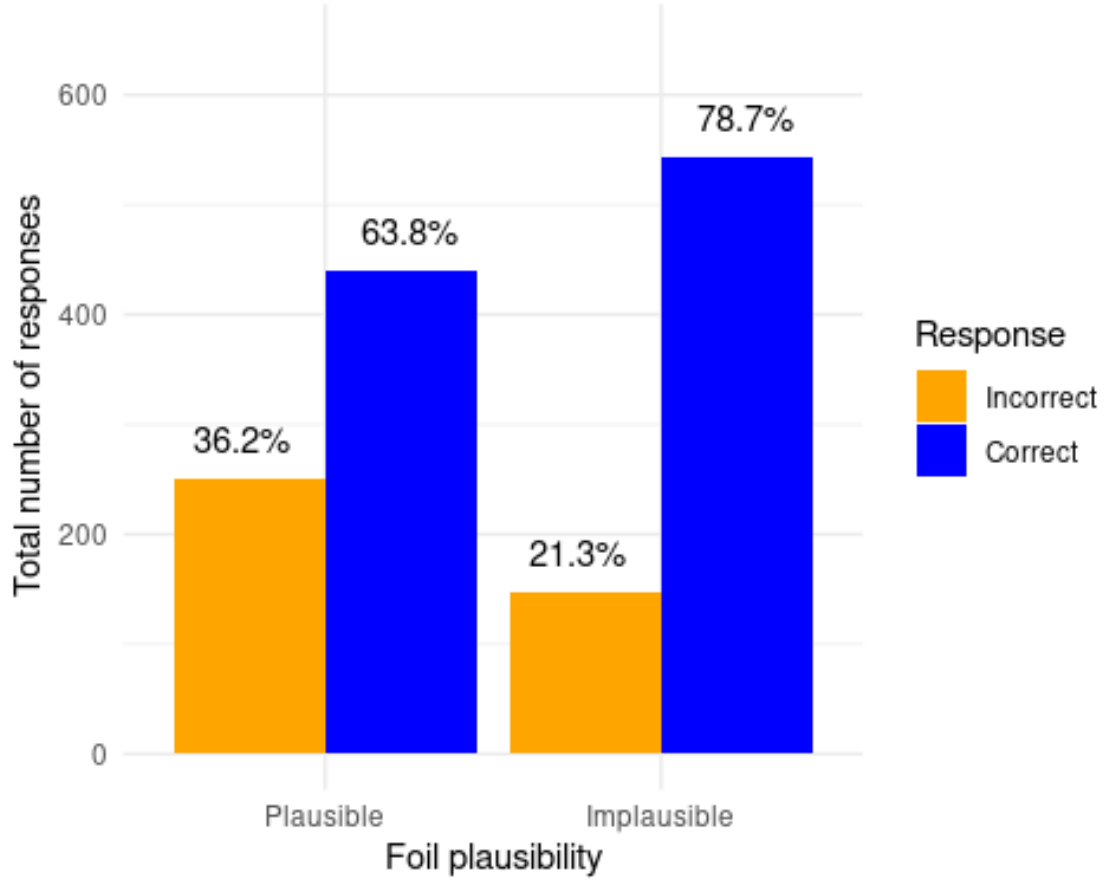

Fig. 7. Number of correct ('Never Seen') and incorrect ('True', 'False', 'No Information') responses to foil statements by plausibility.

[21] glmer(familiarity ~ plausibility + (plausibility | subject)). Figures presenting estimated P(never seen) and P(true |incorrect), respectively, which show the same pattern of results as Figures 7 and 8, can be found in Supplementary Materials 3.

Next, we also checked whether for those mistakenly recognised foils plausibility affected whether they were judged 'True' or 'False'. The mixed-effect logistic regression with the dependent variable 'True' vs 'False' included the predictor variable plausibility and random effects at participant level[22]. There was a significant positive effect of plausibility b = 0.49, SE = 0.15, z = 3.29, p = .001. Hence, participants were more likely to respond with 'True' if the foil was plausible (Figure 8).

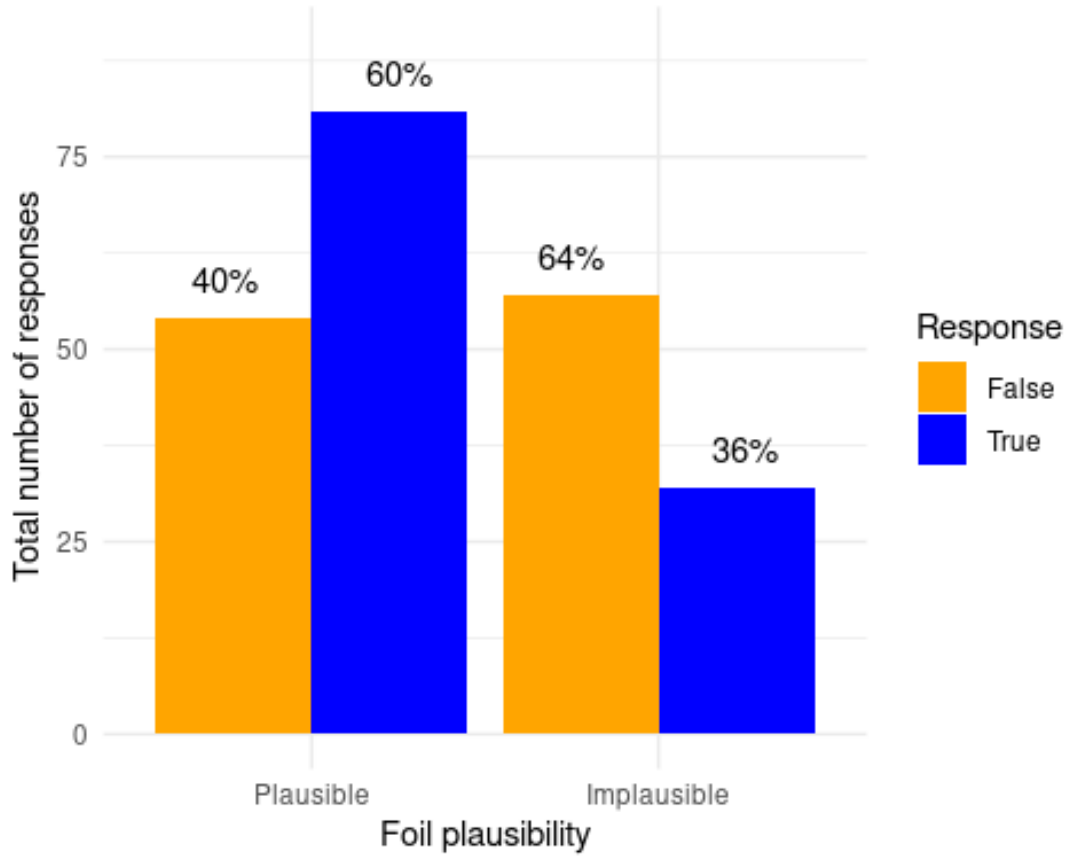

Fig. 8 Number of incorrect 'True' vs 'False' responses to foil statements by plausibility.

Discussion

[22] glmer(responseTrue ~ plausibility + (plausibility | subject)). Significance patterns were identical in a model only including random intercepts at the subject level.

Our key finding is a 'swapping' of Gilbert et al.'s (1990) pattern for reversals with implausible statements: whereas participants tend to misclassify more false statements as true (than true statements as false) when those are plausible (in line with Gilbert et al.'s predictions, [P1]), they misclassify more true statements as false than false statements as true when these statements are implausible [P2], a pattern of results replicated across three studies (Pilot Study SM1, Study 1, Study 2).

As mentioned, plausible statements can be considered the closest to Gilbert et al.'s artificial and hence a-plausible (that is, neither plausible nor implausible) statements. Participants' 'truth bias' with plausible statements can thus be explained by the fact that they are, all in all, more plausible than not, *but also* by the following general hypothesis: in the absence of reasons to doubt, one tends to believe a proposition that is asserted. By no means, however, does that imply that comprehension necessitates acceptance. In fact, the pattern for implausible statements suggests that the so-called 'truth bias' shown by Gilbert et al. should rather be described as a 'plausibility bias' (and thus perhaps not a bias at all, if a bias is defined as a deviation from rationality).

One could be tempted to explain our observed pattern of results by suggesting that participants are just guessing based on plausibility. Results for correct identifications, however, help rule out this explanation: if participants had not learnt anything during the learning phase and answered according to plausibility, one would observe a greater number of correct identifications of plausible true than of implausible true, and of implausible false than of implausible true statements. But instead, participants score as well in correctly identifying implausible true as they do in correctly identifying implausible false statements [P6]. Further, they score as well in correctly identifying implausible true as they do in correctly identifying plausible true statements [P8].

Hence, one crucial aspect of our results is the fact that patterns for correct identifications and reversals are not complementary. Indeed, as just mentioned, implausible true statements are *correctly identified* as often as other types of statements (and more often than plausible false ones). But they are significantly more often *reversed* than implausible false and plausible true statements. This discrepancy between the patterns of correct identifications and reversals for implausible true statements can be explained in terms of surprisingness. Because they are surprising, implausible true statements are more readily remembered during the test phase,

which explains good scores in correct identifications. But when they are not, they are often misclassified as false. That is, among the errors with implausible true statements, there is a significantly high proportion of reversals: they are significantly more often reversed than implausible false and plausible true statements.

The meaning of those results is even clearer when one considers them in terms of *familiarity*. Familiar statements encompass both correct identifications and reversals as statements that participants remember seeing, and being taught 'something' about. *Implausible true statements have the highest familiarity score* (at least numerically), which means that participants remember having seen them and learnt something about their truth-value more often than for any other category of statements. Hence the high proportion of reversals is of prime importance for our purpose: reversals correspond to statements *familiar* to the participants, whose truth-value is falsely reported. That is, participants correctly remember something, namely that they have seen and learnt something about that statement (otherwise, they would go for the 'never seen' or 'no information' answer options), but they then misclassify it. Reversals therefore correspond to cases of *actual misclassification* — of misremembering rather than mere forgetting. Hence, by contrast with cases where participants do not correctly identify a statement because they just do not remember it (or its truth-value), reversals are cases where participants positively believe that a statement is true/false when it is in fact false/true. This reflects that a learning trial might be 'unforgettable' as an experience (e.g., because of its surprisingness) without the truth-value being correctly encoded (otherwise, there would be no reversals at all). Hence, it is possible that surprise also boosts reversals by making trials 'unforgettable'.[23] The two effects of implausibility push in different directions for correct identifications of implausible true statements. On the one hand, the surprise caused by learning that an implausible statement is true might help them to (correctly) remember it, but on the other hand, a statement's implausibility prompts participants to encode it as false. This also means that the two effects might well reinforce each other for reversals of implausible true statements.

[23] As suggested by Nadia Brashier, such effect of surprise on memory could be compared with the hypercorrection effect, whereby errors made with higher confidence are more likely to be corrected with feedback (Butterfield and Metcalfe 2001), a phenomenon likely due to increased attention to surprising feedback (Fazio and Marsh 2009).

Though the ins and outs of our inferences from utilising Gilbert et al.'s (1990) method are complex, the bottom line is simple: we find the opposite of their results on reversals for implausible statements. Crucially, reversals indicate disruption (i.e., the truth-value was not, or not correctly, encoded). But on Gilbert et al.'s account, disrupted statements should always be encoded as true. At test, then, all familiar implausible true statements should be the same: either participants remember them being labelled as true, or they were tagged as true by default. So there really should be no reversals of these statements from the perspective of memory encoding as it is supposed to happen during the learning phase. Plausibility might, of course, exert an effect at recall (during the test phase). But participants correctly recall implausible true statements at a greater rate than any other category. Thus, the effect of plausibility is not merely a response bias. The high number of reversals of implausible true statements indicates that implausible statements are often encoded as false 'per default', which is at odds with Gilbert et al.'s prediction. One might wish to question whether the disruption method can really tackle what occurs at encoding. However, this line of critique leads one to undercut Gilbert et al.'s original (1990) evidence for their account.[24] Thus, one must either accept the method's rationale and the current results which call that account into question or reject the rationale and accept that there is limited experimental evidence for this account.

What does this mean for the nature of belief, and how it is established in the first place? Our results call into question Gilbert's (1991) theory of belief acquisition, whereby the very processes which underlie belief formation result in new statements initially being encoded as true. This matters because Gilbert's account, and the 'truth bias' that comes with it, is often integral to predictions and explanations across many contexts, including misinformation, false beliefs and confirmation bias (Pennycook et al. 2015; Risen, 2016; Kessler et al., 2019), persuasion (Petty et al. 1998; Slater & Rouner, 2002; Green & Brock, 2000), fictional narrative processing (Appel & Richter, 2007; Busselle & Bilandzic, 2008), social networks and online influence (Williams et al., 2017; Marsh & Rajaram, 2019), unintended effects of medical warnings about false claims (Skurnik et al., 2005), the impact of confessions in criminal justice (Appleby & Kassin, 2019), and how children learn from testimony (Harris et al., 2018). Despite

[24] Relatedly, one might argue that memory is a reconstructive process (e.g., Loftus, 1974, 1979; Loftus and Loftus, 1979; Loftus and Hoffman, 1989) and that plausibility may affect reconstruction. But this is just to replace Gilbert's account of what it means to 'have a belief', with an altogether more sophisticated notion, thus rendering his account moot by other means.

the fact that Gilbert's account is invoked in all of these contexts, there has been surprisingly little empirical scrutiny of Gilbert et al.'s (1990) results (the few exceptions are Sperber et al., 2010; Street & Kingstone, 2017; Hasson et al., 2005; Richter et al., 2009; Brashier & Marsh, 2019; Nadarevic & Erdfelder, 2019; see Mercier, 2017, for a review). Our critique of Gilbert et al.'s (1990) findings necessitates a re-evaluation of the evidence and, by extension, of Gilbert and colleagues' theory of belief (e.g., Gilbert, 1991, Gilbert et al., 1993) that is supposed to account for the aforementioned phenomena.

Gilbert's (1991) theory will likely have to be replaced by a view in which the relative plausibility of contents, as well as the reliability of sources of information, play a critical role in the formation of belief. A new account of belief formation in line with these constraints would likely be closer to accepted standards of rationality than Gilbert's account whereby people believe simply by virtue of comprehension. In our study, participants' tendency to accept or reject novel statements seems indeed in line with the prior plausibility of those statements, as revealed by the pretest. In contrast with an approach highlighting the so-called 'cognitive biases' that make us go wrong (see Oaksford and Chater, 1994; Krueger and Funder, 2004; Hahn and Harris, 2014, for critical views), thinking about belief formation from the perspective of 'rational' actors may well prove an important framework within which to understand misinformation and the spread of false beliefs (see also, e.g., Hahn, von Sydow & Merdes, 2018; Desai, Pilditch & Madsen, 2020). These phenomena constitute a serious threat to public health, democratic life, as well as social peace. Detailed attention to the various parameters that (rightly) influence acceptance and memory encoding (including plausibility) seems a more promising perspective to study and further counteract them than positing a (seemingly wrong) fundamental, and rather irrational, bias to believe *any* content.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 660187, and from the *Institut universitaire de France*.

The authors would like to thank David A. Lagnado for his key contribution to the original design of the study, Richard P. Cooper for helpful advice on distraction tasks, Maarten

Speekenbrink for invaluable insight into statistical models, as well as Nadia Brashier, Ryan McKay, and an anonymous reviewer for helpful comments on an earlier draft of this manuscript.

Bibliography

Appel, M., Richter, T. (2007). Persuasive effects of fictional narratives increase over time. *Media Psychology, 10*, 113-114.

Appleby, S., Kassin, S. (2019). When Self-Report Trumps Science: Effects of Confessions, DNA, and Prosecutorial Theories on Perceptions of Guilt. *Psychology, Public Policy, and Law*. In press.

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

Brashier, N., Marsh, E. (2019). Judging Truth. *Annual Review of Psychology, 71*(1):1-17.

Busselle, R., Bilandzic, H. (2008). Fictionality and Perceived Realism in Experiencing Stories: A Model of Narrative Comprehension and Engagement. *Communication theory, 18,* 255–280.

Butterfield, B., Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1491– 1494.

Clément, F., Koenig, M. A., Harris, P. (2004). The ontogenesis of trust. *Mind & Language, 19,* 360–379. http://dx.doi.org/10.1111/j.0268-1064 .2004.00263.x

Coady, C.A.J. (1973). Testimony and Observation. *American Philosophical Quarterly, 10*, 149-155.

Davidson, D. (1967). Truth and meaning. *Synthese, 17*(1), 304-323.

Desai, S. A. C., Pilditch, T. D., & Madsen, J. K. (2020). The rational continued influence of misinformation. *Cognition*, *205*, 104453.

Dummett, M. (1959). Truth. *Proceedings of the Aristotelian Society, 59*(1), 141-162.

Dummett, M. (1976). What is a theory of meaning? (II). In G. Evans & J. McDowell (Eds.), *Truth and Meaning: Essays in Semantics* (pp. 67-137). Oxford: Clarendon Press.

Ebert P., Smith M. (eds.). (2012). Dialectica. *Special Issue: Outright Belief and Degree of Belief*, *66*(3).

Fazio, L. K., Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review,* 16, 88–92. doi:10.3758/PBR.16.1.88

Fazio, L. K., Rand, D. G., Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic bulletin & review*, *26*(5), 1705–1710. https://doi.org/10.3758/s13423-019-01651-4

Fiske, S. T. (2018). Stereotype Content: Warmth and Competence Endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

Gelfert, A. (2014). *A Critical Introduction to Testimony*. Bloomsbury Academic.

Gilbert, D.T., Krull, D.S., Malone, P.S. (1990). Unbelieving the unbelievable. Some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59*(4), 601-613.

Gilbert, D.T. (1991). How mental systems believe. *American Psychologist, 46*(2), 107-119.

Gilbert, D.T., Tafarodi, R.W., Malone, P.S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology, 65*(2), 221-233.

Green, M., Brock, T. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*,*79*(5), 701-721.

Hahn, U., Harris, A.J.L., Corner, A. (2009). Argument Content and Argument Source: An Exploration. *Informal Logic.* 29. 337-367. 10.22329/il.v29i4.2903.

Hahn, U., Harris, A.J.L. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of Learning and Motivation, 61*, 41-102.

Hahn, U., Merdes, C., von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science, 10*(4), 660-678.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Cambridge, MA: Belknap Press/Harvard University Press. http:// dx.doi.org/10.4159/harvard.9780674065192

Harris, P. L., & Lane, J. D. (2013). Infants understand how testimony works. *Topoi, 33,* 443.

Harris, P., Koennig, M., Corriveau, K., Jaswal, K. (2018). Cognitive Foundations of Learning from Testimony. *Annual Review of Psychology 69*, 251-273.

Hasson, U., Simmons, J.P., Todorov, A. (2005). Believe it or not. On the possibility of suspending belief. *Psychological Science, 16*(7), 566-571.

Higginbotham, J. (1986). Linguistic theory and Davidson's program in semantics. In E. LePore (Ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson* (pp. 29-48). Cambridge: Blackwell.

Higginbotham, J. (1989). Elucidations of Meaning. *Linguistics and Philosophy, 12*(4), 465-517.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist, 58*(9), 697-720.

Kessler, E.D., Braash, J.L.G., Kardash, C.M. (2019). Individual Differences in Revising (and Maintaining) Accurate and Inaccurate Beliefs About Childhood Vaccinations. *Discourse Processes: A Multidisciplinary Journal, 56*(5-6), 415-428.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27, 313–327.

Lackey, J., Sosa, E., Eds. (2006). *The Epistemology of Testimony*. Oxford: Oxford University Press.

Lewandowsky, S., Ecker, U.K.H., Seifert, C.M., Schwarz, N., Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, Supplement, 13*(3), 106-131.

Loftus, E.F. (1974). Reconstructing memory: The incredible eyewitness. *Psychology Today, 8,* 116–119.

Loftus, E.F. (1979). The malleability of human memory. *American Scientist, 67*, 312–320.

Loftus, G.R., Loftus, E.F. (1976). *Human Memory: The Processing of Information.* Hillsdale, NJ: Erlbaum Associates.

Loftus, E.F., Hoffman, H.G. (1989). Misinformation and memory: The creation of memory. *Journal of Experimental Psychology: General, 118,*100–104.

Mandelbaum, E. (2014). Thinking is believing. *Inquiry: An Interdisciplinary Journal of Philosophy, 51*(1), 55-96.

Marsh, E., Rajaram, S. (2019). The Digital Expansion of the Mind: Implications of Internet Usage for Memory and Cognition. *Journal of Applied Research in Memory and Cognition, 8*(1), 1-14.

Mascaro, O., Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition, 112,* 367–380. http://dx.doi.org/10.1016/j.cognition.2009.05.012

Mercier, H. (2017). How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology, 21*(2), 103-122.

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology, 49,* 404–418. http://dx.doi.org/10.1037/a0029500

Nadarevic, L., Erdfelder, E. (2019). More evidence against the Spinozan model: Cognitive load diminishes memory for "true" feedback. *Memory & Cognition, 47* (7), 1386–1400.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.

Pennycook, G., Cheyne, J.A., Barr, N., Koehler, D.J., Fugelsang, J.A. (2015). On the reception and detection of pseudo-profound bullshit? *Judgment and Decision-Making, 10*(6), 549-563.

Petty, R. E., Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 323-390). New York, NY, US: McGraw-Hill.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Richter, T., Schroeder, S., Wöhrmann, B. (2009). You don't have to believe everything you read: background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology, 96*(3), 538-558.

Risen, J. (2016). Believing What We Do Not Believe: Acquiescence to Superstitious Beliefs and Other Powerful Intuitions. *Psychological Review, 123*(2), 182–207.

Singmann, H., Bolker, B., Westfall, J. & Aust, F (2016). afex: Analysis of Factorial Experiments. R package version 0.16-1. https://CRAN.R-project.org/package=afex

Skurnik, I., Yoon, C., Park, D.C., Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Memory and Language, 31*, 713-724.

Slater, M.D., Rouner, D. (2002). Entertainment-Education and Elaboration Likelihood: Understanding the Processing of Narrative Persuasion. *Communication Theory, 12*(2), 173-191.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., Wilson, D. (2010). Epistemic vigilance. *Mind and Language, 25*(4), 359-393.

Street, C.N.H., Kingstone, A. (2017). Aligning Spinoza with Descartes: An informed Cartesian account of the truth bias. *British Journal of Psychology, 108*(3), 453-466.

Williams, E., Beardmore, A., Joinson, A., (2017). Individual differences in susceptibility to online influence: A theoretical review. *Computers in Human Behavior,* 72, 412-421.

Supplementary materials 1: Pilot study (SM1).

In an initial, exploratory study, we aimed at replicating Gilbert's 'Hopi language experiment' as closely as possible, while introducing a plausibility variable, and (to this end) using a distinct kind of propositions as materials. In this first study, we also manipulated cognitive load.

Beside this, the procedure was essentially the same as in our confirmatory study.

Method

Participants

105 participants were recruited and paid via Amazon Mechanical Turk (M_{age}=39 [21-70], SD=11,1). They were paid 3.75$ for this ~25-minute task.

Design

We employed a 2x2x2 repeated measures design. The independent variables were plausibility (Plausible or Implausible), veracity (True or False), and disruption (Not disrupted or Disrupted).

Each of the 24 critical statements fell into one of the following 8 categories: Non-disrupted plausible true, Disrupted plausible true, Non-disrupted plausible false, Disrupted plausible false, Non-disrupted implausible true, Disrupted implausible true, Non-disrupted implausible false, Disrupted implausible false.

Beside this, the design and materials were essentially the same as in our confirmatory study.

Procedure

The only difference with our confirmatory study was manipulation of cognitive load during learning phase.[25]

The 48 statements were grouped into eight blocks of six statements. Half of the blocks were 'disrupted' blocks, half of the blocks were 'not disrupted' ones, following an alternating pattern (1st block was disrupted, 2nd was not, … 7th block was disrupted, 8th was not). The block design was not apparent to the participants, who saw the 48 statements successively.

In the not disrupted blocks, each statement appeared for 8 seconds, followed by a 2 seconds blank screen. Immediately after the blank screen, participants saw either another blank screen, or a signal word TRUE or FALSE, for 1 second. After another 1 second blank screen, the next statement appeared.

In the disrupted blocks, each statement appeared for 8 seconds, followed by a 2 second blank screen, itself followed by a voice from the computer giving a random sequence of 5 digits (*e.g.* 6-0-7-4-5). Immediately after hearing the sequence, participants saw either a blank screen, or a signal word TRUE or FALSE, for 1 second, after which a sentence on the computer screen asking the participant to type the sequence of digits she had just heard. Participants were allowed 8 seconds to type the sequence. Then the next statement appeared.

Timings were chosen to replicate Gilbert et al.'s setup as closely as possible.

The 1st and last (8th) blocks contained only buffers (against primacy and recency effects, like in Gilbert et al.'s setup), in a randomised order. Like the statements in the other 6 blocks, some buffers were followed by a signal word (TRUE or FALSE) some by a blank screen. The 1st block was disrupted, the 8th block was not.

From the 2nd to the 7th (penultimate) blocks, each block contained, in a randomised order, four critical statements randomly drawn from each of the four categories, plausible true, plausible false, implausible true, implausible false, and two "fillers" (statements with no truth value tag after them).

[25] We had to run several pilots until we found a disruption task that impaired performance, and thus functioned as a disruption task. One tentative hypothesis about why Gilbert's task did not have any distraction effect on our participants could be that our materials, being about the real world, attracted greater attention than Gilbert et al.'s (1990) 'Hopi words' definitions, such that a stronger disruption task was needed. Also, there were effects of veracity and of plausibility in the non-disrupted conditions also; given that those effects are what we are interested in, and that cognitive load was mostly aimed at making them more salient, we set out the study as soon as we found a disruption task that had a main effect on correct identifications (although the effect was not very strong, and there was no effect on reversals).

The test phase was the same as in our confirmatory study.

Results

Results showed non-complementary patterns for correct identification and reversals. Since reversals are our key data, we present them first.

Reversals

Our most important result was a statistical interaction between plausibility and veracity overall (for both disrupted and non-disrupted trials). As appears in Figure 1, there were significantly more reversals when truth value was in opposition with plausibility (i.e., for plausible false and implausible true statements), in both (disrupted and non-disrupted) conditions.

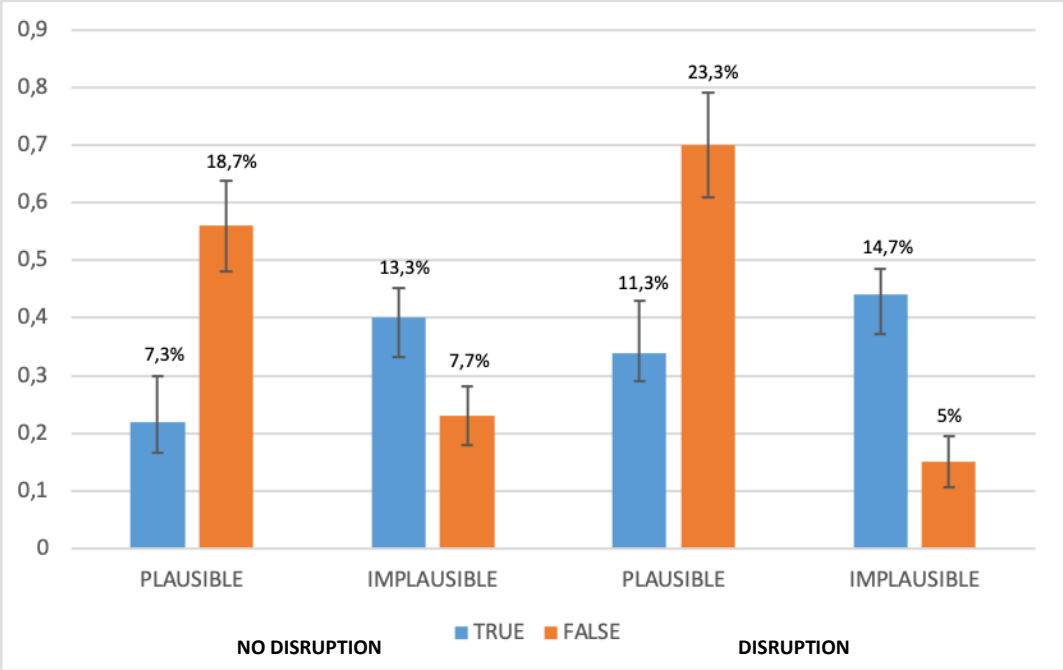

Fig. 1. Reversals, in the 'no disruption' (left) and 'disruption' (right) conditions (absolute number shown on x-axis and written as a percentage of items). Error bars are plus and minus 1 standard error

We ran a 2(disruption)x2(veracity)x2(plausibility) within subject repeated measures ANOVA. Whilst there was no three way interaction, $F(1,104)=.665$, $p=.417$, $\eta_p^2=.006$, the interaction of veracity and plausibility was significant, $F(1, 104)=28.780$, $p<.001$, $\eta_p^2=.217$. We found a main effect of plausibility, $F(1, 104)=23.184$, $p=<.001$, $\eta_p^2=.182$, but no main effect of disruption, $F(1,104)=2.736$, $p=.101$, $\eta_p^2=.026$, nor of veracity, $F(1, 104)=2.328$, $p=.130$, $\eta_p^2=.022$. There was no interaction between disruption and veracity, $F(1,104)=.432$, $p=.512$, $\eta_p^2=.004$, and the interaction between disruption and plausibility was borderline significant, $F=3.671$, $p=.058$, $\eta_p^2=.032$.

Since disruption did not have any significant effect on reversals (note however that it did so on correct identifications)[26], we ignore it here, and explore the interaction between plausibility and veracity (overall). Overall (independent from whether trials were disrupted or not), participants mistake more false statements as true than true as false when those are plausible (which corresponds to Gilbert's pattern)[27], but they mistake more true statements as false (than false as true) when those are implausible (swapping of Gilbert's pattern). Participants tend to misclassify true statements as false more often when those are implausible than when they are plausible, and to misclassify more false statements as true when those are plausible than when they are implausible.

These observations were confirmed by one-way ANOVAs, which revealed significant effects of veracity on plausible statements, $F(1, 104)=20.100$, $p<.001$, $\eta_p^2=.162$, as well as on implausible statements, $F(1, 104)=17.547$, $p<.001$, $\eta_p^2=.144$. They also revealed significant effects of plausibility on true statements, $F(1,104)=5.908$, $p=.017$, $\eta_p^2=0.54$, as well as on false

[26] Four 2-way ANOVAs were run, to investigate the interaction of disruption and veracity on plausible / implausible statements respectively, as well as the interaction of disruption and plausibility on true / false statements respectively, which revealed no significant interaction, except a small one for disruption x plausibility for false statements, F=4.095, p=.046, $\eta_p^2=.038$. We therefore concentrate on the respective effects of plausibility (on true / false statements), and of veracity (on implausible / plausible statements) that such analyses revealed.

[27] Replicating Gilbert's results for plausible statements would imply finding an interaction between veracity and disruption, which we didn't (F=.28, p=.868, $\eta_p^2<.001$), probably because disruption effect is weak. There was however a small effect of disruption on plausible statements (F=6.137, p=.015, $\eta_p^2=.056$).

statements, $F(1,104)=42.277$, $p<.001$, $\eta_p^2=.289$. Let us now consider results for correct identifications.

Correct identifications

As already mentioned, the pattern for correct identifications was not simply the complement to the one for reversals (see figure 2). Most notably, the interaction between veracity and plausibility was not significant (see details below), and there was no effect of veracity on implausible statements (participants didn't perform better — nor worse — in identifying implausible false than implausible true statements). On the other hand, plausible true statements were significantly better identified than plausible false ones. In brief, whilst for plausible statements the pattern was the complement to the one for reversals (and thus in line with Gilbert's pattern, as we predicted), it was not the case for implausible ones. It is also worth noting that participants performed above chance in all conditions when not disrupted.

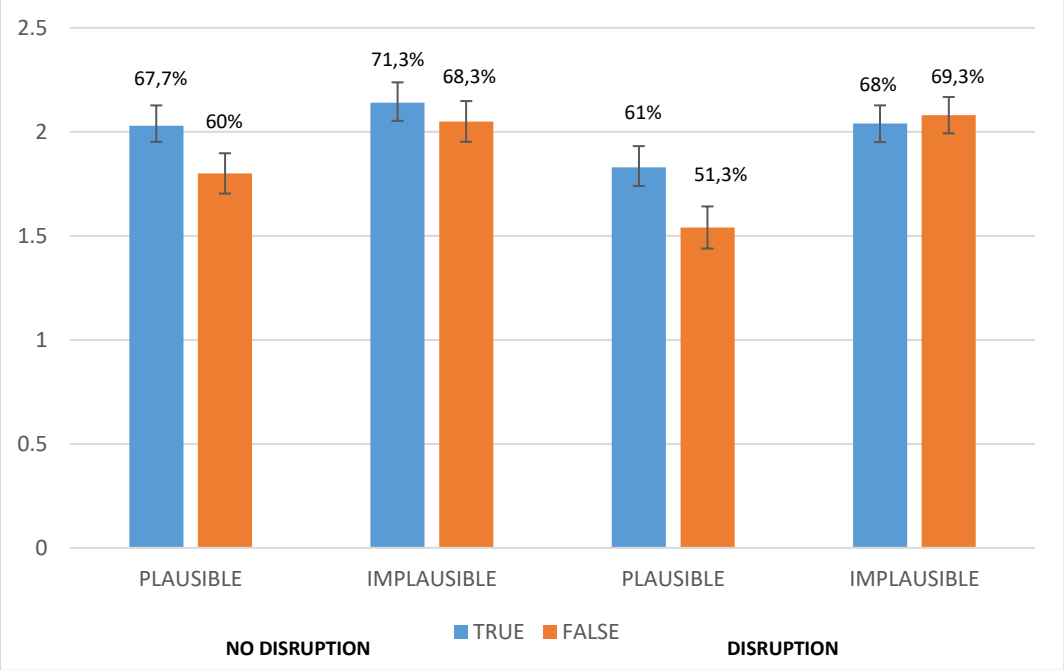

Fig. 2. Correct identifications, in the 'no disruption' (left) and 'disruption' (right) conditions (absolute number shown on x-axis and written as a percentage of items). Error bars are plus and minus 1 standard error.

A 2(disruption)x2(veracity)x2(plausibility) repeated measures ANOVA revealed no three-way statistical interaction, $F(1,104)=.802$, $p=.373$, $\eta_p^2=.008$. Neither was any two-way interaction significant.[28] but there were main effects of disruption, $F(1,104)=5.210$, $p=.024$, $\eta_p^2=.048$, of plausibility, $F(1,104)=23.225$, $p<.001$, $\eta_p^2=.183$, and of veracity, $F(1,104)=6.237$, $p=.014$, $\eta_p^2=.057$.

Just like for reversals (see footnote 2), four 2-way ANOVAs were run so as to investigate interaction of disruption with veracity (on plausible / implausible statements respectively) and of disruption with plausibility (on true / false statements respectively), which revealed no significant interaction[29]. Disruption had an effect on correct identification of plausible statements (in line with Gilbert's results), F=7.761, p=.006, η_p^2= 0.69, but not of implausible ones, F=.272, p=.603, $\eta_p^2=.003$.

Although interaction between plausibility and veracity was not significant, $F(1,104)=2.769$, $p=.099$, $\eta_p^2=.026$), one way ANOVAS investigating the effects of veracity and of plausibility on plausible / implausible, and true / false statements respectively, were conducted.

We found a significant effect of veracity on plausible statements, $F(1,104)=7.517$, $p=.007$, $\eta_p^2=.067$, participants scoring better in correct identification of plausible true than of plausible false statements, but no effect of veracity on implausible statements, F=.114, p=.737, $\eta_p^2 = .001$).

The effect of plausibility on true statements (participants scoring slightly better for implausible true than for plausible true statements) was not significant, $F(1,104) = 3.502$, $p = .064$, $\eta_p^2 = .033$. However, we found a significant effect of plausibility on false statements,

[28] There was no interaction between disruption and veracity, $F(1,104) = .105$, $p=.746$, $\eta_p^2 = .001$. Neither interaction between plausibility and veracity, F = 2.769, p = .099, $\eta_p^2 = .026$, nor interaction between disruption and plausibility was significant F(1,104) = 3.454, $p=.066$, $\eta_p^2 = .032$.

[29] Interaction disruption x plausibility for false statements was borderline significant.

F(1,104) = 17.905, p<.001, η_p^2 = .147, participants scoring significantly better for implausible false than for plausible false statements.

Supplementary Materials 2:

Materials (full list of statements) for Studies 1 & 2

STUDY 1

Critical statements:

	TRUE	FALSE
PLAUSIBLE	In Luxembourg, the minimum age of criminal responsibility (i.e. the minimum age at which a person can be prosecuted for a crime) is 18.	In Portugal, the minimum age of criminal responsibility (i.e. the minimum age at which a person can be prosecuted for a crime) is 18.
	In Singapore, homosexual relations are prohibited.	In Taiwan, homosexual relations are prohibited.
	The lion is the second-largest living cat.	The tiger is the second-largest living cat.
	In Denmark, drivers are supposed to drive their vehicles with their headlights on during the day as well.	In Norway, drivers are supposed to drive their vehicles with their headlights on during the day as well.
	Spain is the third biggest wine producer in the world.	The USA are the third biggest wine producer in the world.
	Homosexuals can be sentenced to d Sudan.	Homosexuals can be sentenced to death in Uganda.
	Same-sex marriage is legally recognized in New Zealand.	Same-sex marriage is legally recognized in Australia.
	A leopard can live up to 17 years.	A cheetah can live up to 17 years.

	Homosexuality is illegal in Guyana. In Singapore, dropping trash on the ground can require you to pay a $1000 fine. In the wild, lionesses live 10 to 14 years. Italy was the world's biggest wine producer in 2015 ahead of France.	Homosexuality is illegal in Suriname. In Switzerland, dropping trash on the ground can require you to pay a $1000 fine. In the wild, male lions live 10 to 14 years. France was the world's biggest wine producer in 2015 ahead of Italy.
IMPLAUSIBLE	In Milan (Italy), it is a legal requirement to smile at all times, except during funerals or hospital visits. Flushing the toilet after 10pm is illegal in Switzerland. In France, you can marry a dead person. In England, it is illegal for women to eat chocolate on public transport. In Florida, the law forbids single women to parachute on Sundays. It is illegal to make funny faces at	In Padua (Italy), it is a legal requirement to smile at all times, except during funerals or hospital visits. Flushing the toilet after 10pm is illegal in Luxembourg. In Spain, you can marry a dead person. In Ireland, it is illegal for women to eat chocolate on public transport. In South Carolina, the law forbids single women to parachute on Sundays. It is illegal to make funny faces at a

	a dog in Oklahoma.	dog in Missouri.
	In Florida, it is illegal to fart in a public space after 6pm on Thursdays.	In South Carolina, it is illegal to fart in a public space after 6pm on Thursdays.
	Scientists can turn peanut butter into diamonds.	Scientists can turn almond oil into diamonds.
	Herring communicate through farts.	Mackerel communicate through farts.
	Female kangaroos have three vaginas.	Female wallaby have three vaginas.
	In England, mince pies cannot be eaten on Christmas Day.	In Scotland, mince pies cannot be eaten on Christmas Day.
	In Bangkok, police officers who commit minor transgressions will have to wear a bright pink Hello Kitty armband for several days.	In Rangoon, police officers who c minor transgressions will have to v bright pink Hello Kitty armband for s days.

Fillers were the same for all participants (presented in a randomized order, only during the learning phase):

In France, it was against the law for women to wear trousers until 2013.

In Pennsylvania (US), there is a law that prohibits the sale of condoms from vending machines.

In India, cheating in an academic exam can bring you to jail.

It is forbidden to drive with sandals or flip flops in Spain.

In Wisconsin, USA, it is against the law to hang female and male underwear together on the same washing line.

In Sweden, it is illegal to start your car without first checking to see if there are any children sleeping under it.

In Jesolo, near Venice (Italy), it is illegal to build sand castles on the beach

India is the largest exporter of lentils in the world.

In Ottawa (Canada), it is illegal to eat ice cream on Bank Street on Sundays.

In France, 70% of music on radio from 8am to 8pm must be French music.

In Arkansas, it is against the law to have a sleeping donkey in your bathtub after 7pm

The fertility rate (number of children per woman in childbearing age) in 2015 in Taiwan was less than 1

Buffers were the same for all participants (6 at the beginning and 6 at the end of the learning phase, in a randomized order, some indicated as true, some as false, and some with blank screen):

In Paris (France), feeding pigeons is forbidden, and can lead to fines up to 450 euros. BLANK SCREEN

Ottawa, Canada, is the second coldest capital in the world. TRUE

In Rome (Italy), it is forbidden to eat in the historic centre. BLANK SCREEN

The speed limit for automobiles on expressways in New Caledonia is 110. TRUE

In Alabama, sodomy is illegal. TRUE

The longest pregnancy in humans on record is 374 days (one year and 9 days). BLANK SCREEN

Chewing gum is illegal in Taiwan. FALSE

Rats and human DNA are 97.5% similar. FALSE

The gestation period of the guinea-pig is 6 weeks. BLANK SCREEN

China is the second largest country in the world. FALSE

The infant mortality rate (death of children under one year of age) in 2015 in Switzerland was 4 per 1000. BLANK SCREEN

There are more saunas than cars in Finland. BLANK SCREEN

Foils (only used in the test phase, same seen by all participants, *mutatis mutandis* — when participants saw a statement about South Florida, the "Florida" version was used, same with Padua / Milan, Toronto / Ottawa)

It is illegal to be overweight in Japan.

In Switzerland, you need a licence to drive a bicycle.

Running of petrol is illegal on Sweden's motorways.

In Canada, by law, one out of every five songs on the radio must be sung by a Canadian.

It is forbidden to eat in the street in the historic centre in Milan / Padua (Italy).

It is illegal not to flush a public toilet in Singapore.

Chewing gum is illegal in Singapore.

In Ottawa / Toronto (Canada), it is illegal to eat chocolate on Bank Street on Sundays.

In France, is illegal to name a pig Napoleon.

In South Carolina / Florida, it was illegal for unmarried couples to live together until April 2016.

The gestation period for a hamster is 6 weeks.

Homosexual relations are prohibited in Nigeria.

STUDY 2 (replication study)

Critical statements (those not used in experiment 1 appear in bold characters):

	TRUE	FALSE
PLAUSIBLE	In Luxembourg, the minimum age of criminal responsibility (i.e., the minimum age at which a person	In Portugal, the minimum age of criminal responsibility (i.e., the minimum age at which a person can

	can be prosecuted for a crime) is 18.	be prosecuted for a crime) is 18.
	'Taaqa' means 'man' in Hopi language.	**'Taawa' means 'man' in Hopi language.**
	Same-sex marriage is legally recognized in Finland.	**Same-sex marriage is legally recognized in Lithuania.**
	The speed limit in urban areas in Costa Rica is 40.	**The speed limit in urban areas in Ecuador is 40.**
	Cellophane tape was invented by an American.	**Cellophane tape was invented by a Canadian.**
	Running out of petrol is illegal on Germany's motorways, and so is walking along it.	**Running out of petrol is illegal on France's motorways, and so is walking along it.**
	'Gahni' means 'house' in Shoshone language.	**'Gahni' means 'water' in Shoshone language.**
	It was a pair of Canadians who created the game Trivial Pursuit.	**It was a pair of Americans who created the game Trivial Pursuit.**
	Homosexuality is illegal in Guyana.	Homosexuality is illegal in Suriname.
	In Singapore, dropping trash on the ground can require you to pay a $1000 fine.	In Switzerland, dropping trash on the ground can require you to pay a $1000 fine.
	It is forbidden to drive with sandals or flip flops in Spain	**It is forbidden to drive with sandals or flip flops in France.**
	Italy was the world's biggest wine producer in 2015 ahead of France.	France was the world's biggest wine producer in 2015 ahead of Italy.

IMPLAUSIBLE	In Milan (Italy), it is a legal requirement to smile at all times, except during funerals or hospital visits.	In Padua (Italy), it is a legal requirement to smile at all times, except during funerals or hospital visits.
	In Ottawa (Canada), it is illegal to eat ice cream on Bank Street on Sundays.	**In Toronto (Canada), it is illegal to eat ice cream on Bank Street on Sundays.**
	In France, it was against the law for women to wear trousers until 2013.	**In Italy, it was against the law for women to wear trousers until 2013.**
	In England, it is illegal for women to eat chocolate on public transport.	In Ireland, it is illegal for women to eat chocolate on public transport.
	In Florida, the law forbids single women to parachute on Sundays.	In South Carolina, the law forbids single women to parachute on Sundays.
	It is illegal to make funny faces at a Oklahoma.	It is illegal to make funny faces at a Missouri.
	In Florida, it is illegal to fart in a public space after 6pm on Thursdays.	In South Carolina, it is illegal to fart in a public space after 6pm on Thursdays.
	Scientists can turn peanut butter into diamonds.	Scientists can turn almond oil into diamonds.
	Oxford University is older than the Aztecs.	**Salamanca University is older than the Aztecs.**

	The longest pregnancy in humans on record is 375 days (one year and 10 days). **In France, you can marry a dead person.** In Bangkok, police officers who commit minor transgressions will have to wear a bright pink Hello Kitty armband for several days.	**The longest pregnancy in humans on record is 374 days (one year and 9 days)** **In Spain, you can marry a dead person.** In Rangoon, police officers who commit minor transgressions will have to wear a bright pink Hello Kitty armband for several days.

Fillers were the same for all participants (presented in a randomized order, only during the learning phase; half plausible, half implausible):

<u>Plausible:</u>

The fertility rate (number of children per woman in childbearing age) in 2015 in Taiwan was less than 1.

Marlon Brando was expelled from high school for riding a motorcycle through the highway.

A leopard can live up to 17 years.

In Finland, fines for minor infractions (such as traffic fines) are calculated as a percentage of the offenders' income.

In Arkansas, sodomy is illegal.

In the wild, lionesses live 10 to 14 years.

<u>Implausible:</u>

Iroly is an Island in the Pacific, where daughters can marry their father.

Torre Annunziata, South of Naples, has outlawed miniskirts.

In England, mince pies cannot be eaten on Christmas Day.

Until April 2016, it was illegal for unmarried couples to live together in South Carolina.

Cotton candy was invented by a dentist.

In Eraclea, near Venice (Italy), it is illegal to build sand castles on the beach.

Buffers were the same for all participants (6 at the beginning and 6 at the end of the learning phase, in a randomized order, some indicated as true, some as false, and some with blank screen; in each category, equal number of plausible / implausible).

Plausible (with truth value information)

True:

In Maryland (US), there is a law that prohibits the sale of condoms from vending machines.

False:

The USA is the third biggest wine producer in the world.

India is the largest exporter of lentils in the world.

Implausible (with truth value information)

True:

Flushing the toilet after 10pm is illegal in Switzerland.

Female koalas have three vaginas.

False:

In the 19th century, a French woman from Lyon died aged 199.

Plausible without information (followed by blank screen)

In Switzerland, dropping trash on the ground can require you to pay a $1000 fine.

The infant mortality rate (death of children under one year of age) in 2015 in Israel was 4 per 1000.

In Norway, drivers are supposed to drive vehicles with their headlights on during the day as well.

Implausible without information (followed by blank screen)

You're more likely to get a computer virus from visiting religious sites than porn sites.

Airline pilots are not allowed to grow moustaches.

In Denmark, it is illegal to start your car without first checking to see if there are any children sleeping under it.

Foils (only used in the test phase, same seen by all participants; half plausible, half implausible)

Implausible

In Portugal, it is forbidden to marry someone from a different town. 49

In Paris (France), feeding pigeons is forbidden, and can lead to fines up to 450 euros. 50

In Minnesota, USA, it is against the law to hang female and male underwear together on the same washing line. 51

The speed limit in urban areas in Mexico is 150. 52

Bullfrogs do not sleep. 53

Herrings communicate through farts. 54

Plausible

The infant mortality rate (death of children under one year of age) in 2015 in Switzerland was 4 per 1000.

The fertility rate (number of children per woman in childbearing age) in 2015 in Yemen is 4.

Canada is the second largest country in the world.

The speed limit for automobiles on expressways in New Caledonia is 110.

The gestation period of the guinea-pig is 6 weeks.

In Taiwan, not flushing a public toilet is illegal.

Supplementary Materials 3

Figures SM3-1, SM3-2. SM3-4 and SM3-5 displaying estimates from the mixed effect logistic regressions for P(reversal) and P(correct identification) across the 4 cells of the design in Studies 1 and 2. Figures SM3-3 and SM3-6 displaying P(familiarity) as well as P(correct identification|familiarity). Figure SM3-7 showing effects of plausibility on P(correct) and P(true|incorrect) within foil statements. All plots were created using the *afex* package (Singmann et al., 2016), and represent the estimated marginal means with error bars representing 95% confidence intervals.

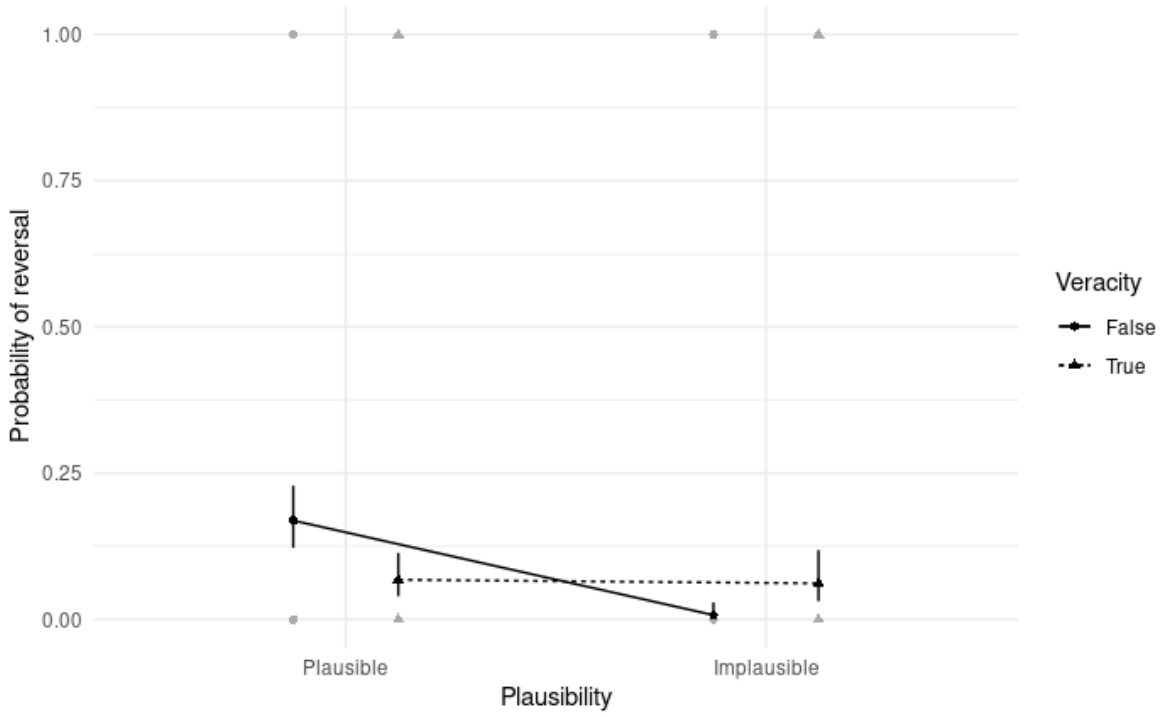

Figure SM3-1. Likelihood of reversal (compared with correct identification, 'no information' and 'never seen' responses) by statement plausibility and veracity (Study 1).

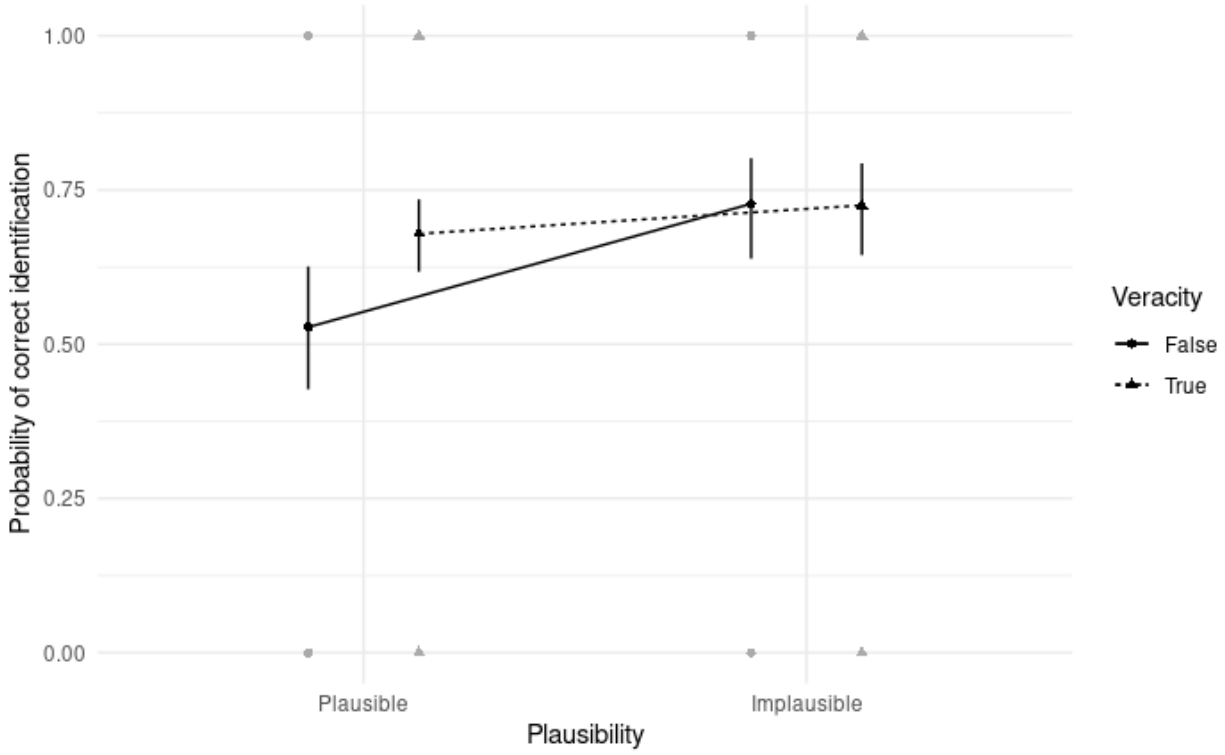

Figure SM3-2. Likelihood of correct identification (compared with reversal, 'no information' and 'never seen' responses) by statement plausibility and veracity (Study 1).

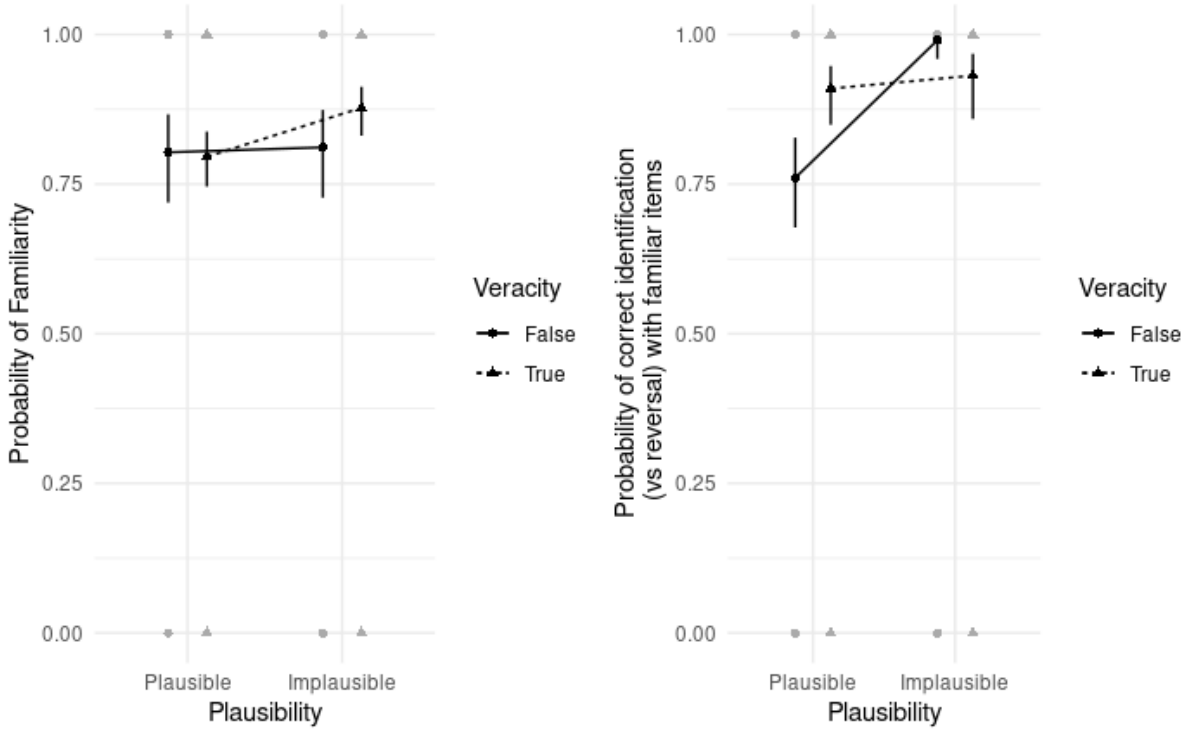

Figure SM3-3. Likelihood of familiarity (left panel) and correct identification within familiar statements (right panel) by plausibility and veracity (Study 1).

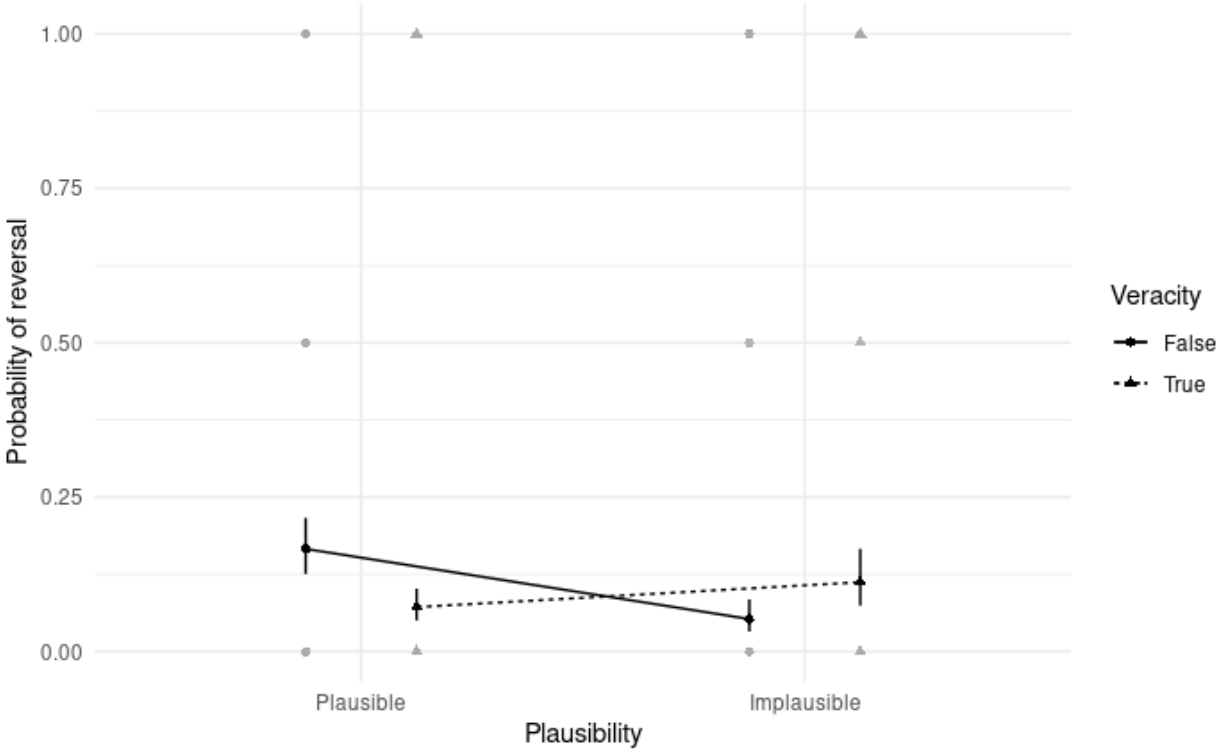

Figure SM3-4. Likelihood of reversal (compared with correct identification, 'no information' and 'never seen' responses) by statement plausibility and veracity (Study 2).

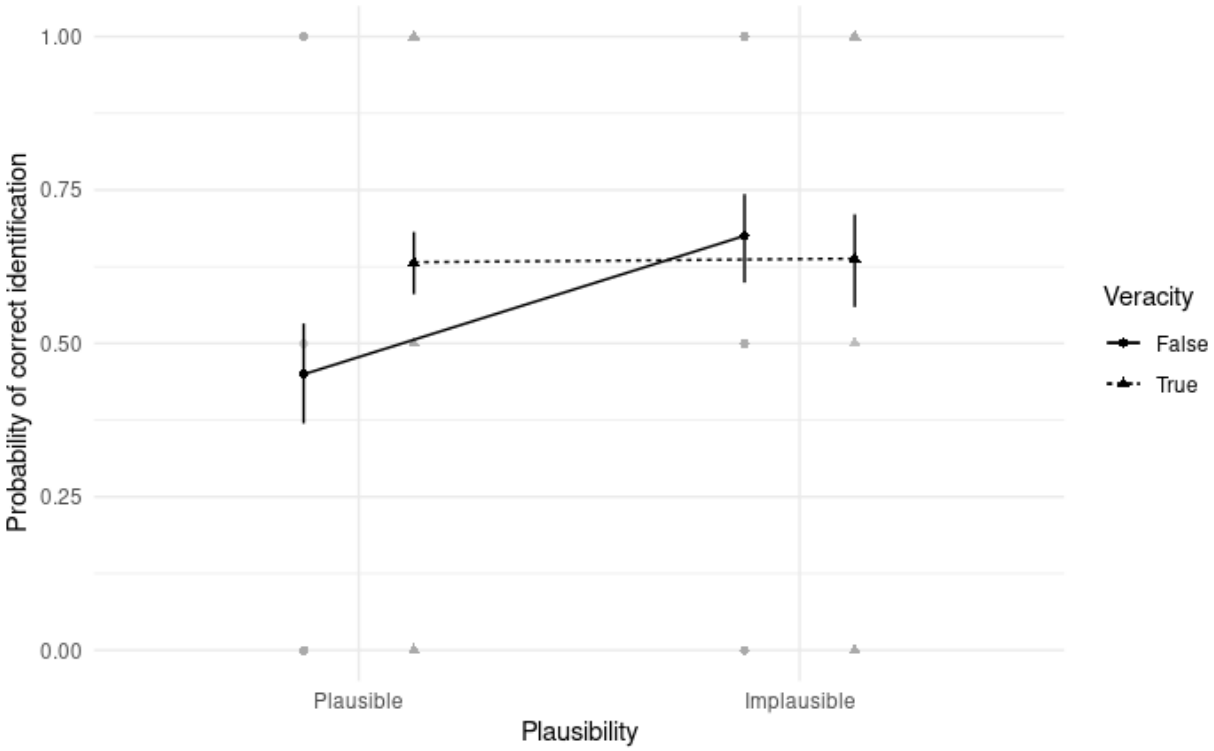

Figure SM3-5. Likelihood of correct identification (compared with reversal, 'no information' and 'never seen' responses) by statement plausibility and veracity (Study 2).

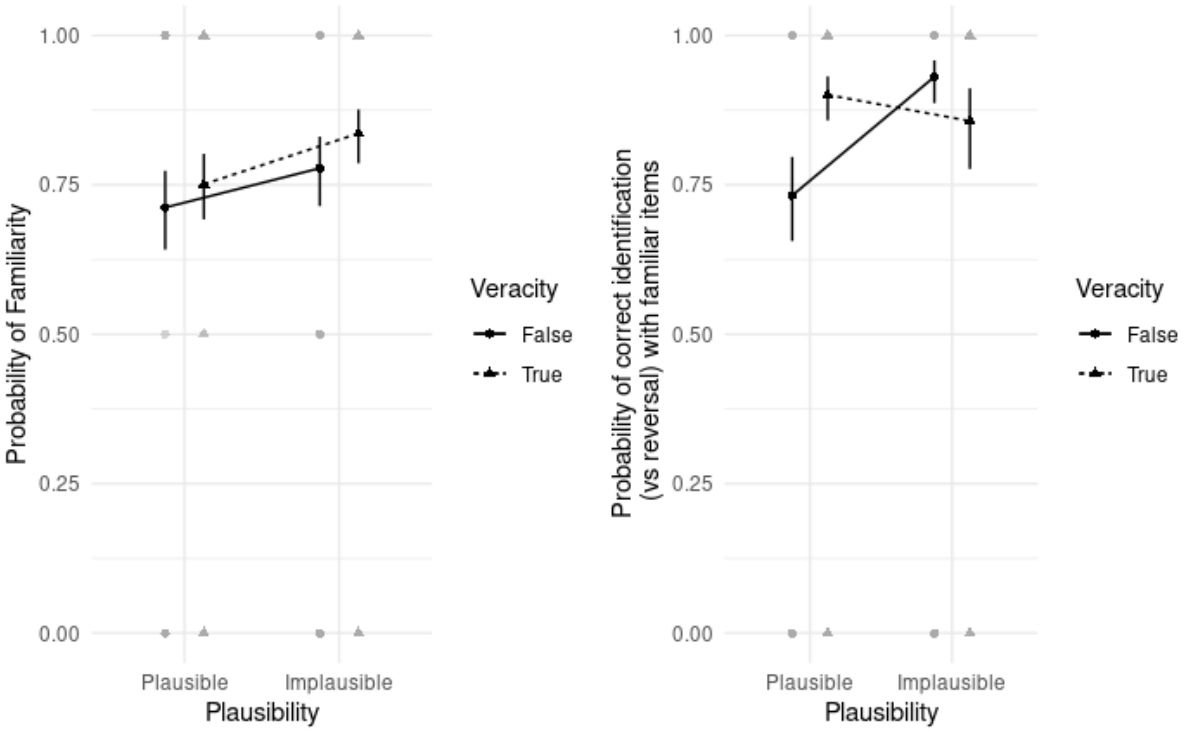

Figure SM3-6. Likelihood of familiarity (left panel) and correct identification within familiar statements (right panel) by plausibility and veracity (Study 2).

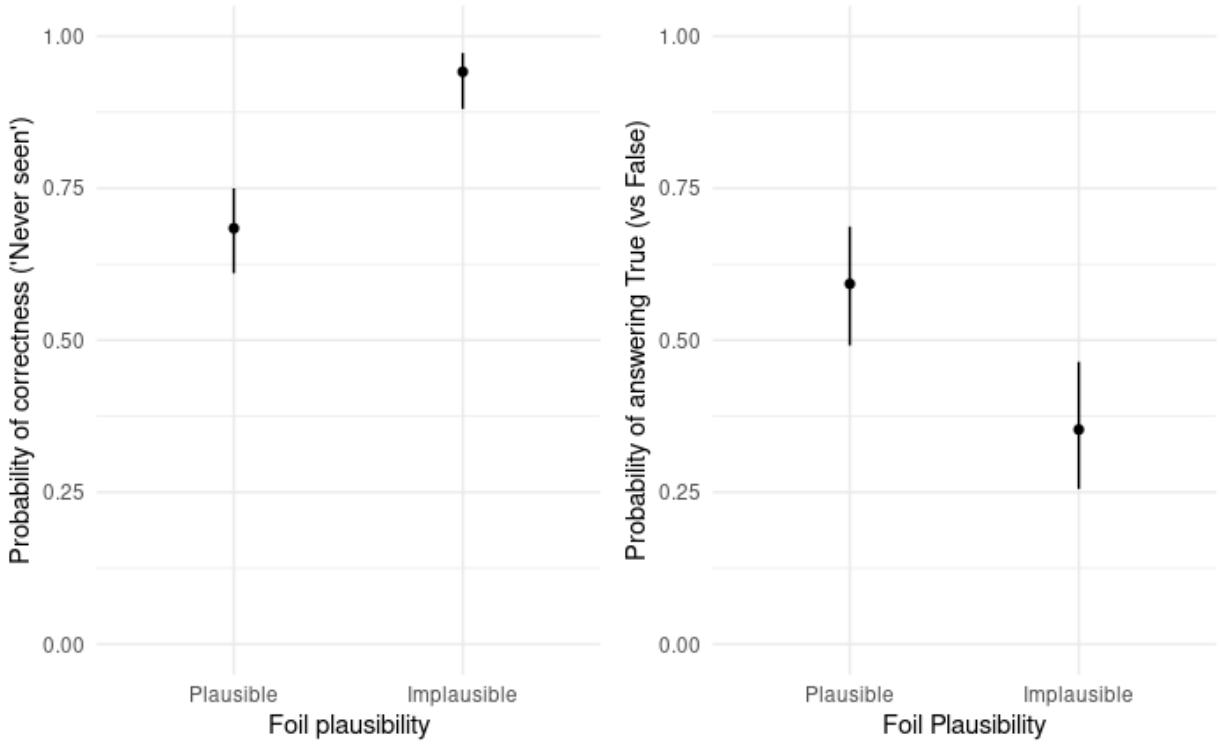

Figure SM3-7. Effects of plausibility on probability of correct responses (left panel) and P(true|incorrect) responses for the foil statements (right panel; Study 2).