Mem Cogn (2013) 41:176–186
DOI 10.3758/s13421-012-0251-z

Spinoza's error: Memory for truth and falsity

Lena Nadarevic · Edgar Erdfelder

Published online: 13 September 2012

Abstract Two theoretical frameworks have been proposed to account for the representation of truth and falsity in human memory: the Cartesian model and the Spinozan model. Both models presume that during information processing a mental representation of the information is stored along with a tag indicating its truth value. However, the two models disagree on the nature of these tags. According to the Cartesian model, true information receives a "true" tag and false information receives a "false" tag. In contrast, the Spinozan model claims that only false information receives a "false" tag, whereas untagged information is automatically accepted as true. To test the Cartesian and Spinozan models, we conducted two source memory experiments in which participants studied true and false trivia statements from three different sources differing in credibility (i.e., presenting 100% true, 50% true and 50% false, or 100% false statements). In Experiment 1, half of the participants were informed about the source credibility prior to the study phase. As compared to a control group, this precue group showed improved source memory for both true and false statements, but not for statements with an uncertain validity status. Moreover, memory did not differ for truth and falsity in the precue group. As Experiment 2 revealed, this finding is replicated even when using a 1-week rather than a 20-min retention interval between study and test phases. The results of both experiments clearly contradict the Spinozan model but can be explained in terms of the Cartesian model.

Keywords Source memory · Memory representation · Truth · Falsity · Multinomial modeling

L. Nadarevic (✉) · E. Erdfelder
Department of Psychology, School of Social Sciences,
University of Mannheim,
68131 Mannheim, Germany
e-mail: nadarevic@psychologie.uni-mannheim.de

E. Erdfelder
e-mail: erdfelder@psychologie.uni-mannheim.de

People steadily encounter a vast amount of information from a variety of sources (e.g., newspaper, television, radio, friends, or colleagues). However, not all information is reliable. Some pieces of information are obviously true (e.g., Paris is the capital of France), whereas others are false (e.g., the Sun revolves around the Earth). For the majority of information, however, the validity status is unknown (e.g., migratory birds fly faster when moving north than when moving south). In these cases, people evaluate whether they consider this information true or false. But what does this evaluation process look like? In the early 1990s, Gilbert and colleagues proposed two theoretical accounts of how people comprehend, evaluate, and represent new information (Gilbert, 1991; Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993). These accounts are based on the basic ideas of the two philosophers René Descartes (1641/1984) and Baruch Spinoza (1677/2006).

The Cartesian model and the Spinozan model

According to the Cartesian model, a mental representation of incoming information is formed and stored in memory during a first processing stage. If a person has sufficient cognitive capacity to assess the validity of the information, its memory representation is then tagged as true or false during a second processing stage. In contrast, if a person lacks capacity to evaluate the information, the memory representation remains unaltered: "I simply refrain from making a judgment in cases where I do not perceive the truth with sufficient clarity and distinctiveness" (Descartes, 1641/1984, p. 41).

Like the Cartesian model, the Spinozan model assumes that a mental representation of new information is stored in memory during a first processing stage. However, unlike Descartes, Spinoza (1677/2006, p. 52) proposed that "Will

and understanding are one and the same." In other words, any information is automatically believed in the first instance (Bennett, 1984). A "false" tag is added to its mental representation only if information is deliberately assessed during a second processing stage and considered to be false. If the information is considered true or if there is not sufficient capacity for a deliberate evaluation process, the representation of the information remains untagged.

The "Hopi" language experiment

To test the Cartesian and Spinozan models empirically, Gilbert et al. (1990) conducted a memory experiment in which words of a fictitious "Hopi" language and their alleged English equivalents were presented as statements (e.g., *A monishna is a star*). In two-thirds of the trials, participants received feedback regarding a statement's validity (true vs. false) immediately after having seen the statement. However, in some trials the processing of this feedback was interrupted by a distractor task.

The rationale behind this procedure was as follows: If the Spinozan model is correct, the interruption of feedback processing should interfere with the encoding of "false" tags. Because the Spinozan model proposes that untagged information is automatically accepted as true, the interruption should result in misremembering ostensibly false statements as true in a subsequent memory test, whereas it should not affect the correct classification of ostensibly true statements. In contrast, if the Cartesian model is correct, distraction should interfere with the encoding of both "false" and "true" tags. As a consequence, memory for truth and falsity should be impaired to a similar degree.

The results of Gilbert et al. (1990) were in line with the Spinozan model. Feedback interruption did not affect correct classifications of ostensibly true statements. However, feedback interruption significantly impaired correct classifications of ostensibly false statements. Moreover, ostensibly false statements were more often misclassified as true when feedback processing had been interrupted. Similar results were observed when feedback processing was disrupted by means of time pressure (Gilbert et al., 1993; Koslow & Beltramini, 2002). Furthermore, the effect did not depend on whether feedback was provided after statement presentation or simultaneously with statement presentation (e.g., Gilbert et al., 1993; Hasson, Simmons, & Todorov, 2005).

The role of information content

The results of the "Hopi" language experiment were conceptually replicated several times—for example, using assertions about an imaginary animal (Gilbert et al., 1990), statements of fictitious crime reports (Gilbert et al., 1993), and product claims (Koslow & Beltramini, 2002). Importantly, however, not all types of information produced the effect observed by Gilbert et al. (1990). Richter, Schroeder, and Wohrmann (2009), for example, replicated the effect only when using general knowledge statements for which participants had no or weak background beliefs (e.g., *Toothpaste contains sulfur*), but not when using statements for which participants had strong background beliefs (e.g., *Soft soap is edible)*. The authors concluded that relevant background knowledge induces fast and efficient validation processes and thus prevents people from automatically accepting everything they comprehend as being true.

Moreover, Hasson et al. (2005) proposed that statement content may also play a crucial role with respect to the memory representation of falsity. According to their account, the Spinozan tagging system applies only to statements that are uninformative when being false (e.g., *A monishna is a star*). However, when a false statement is informative, "then the false statement may be represented in terms of what its falsity implies or suggests" (Hasson et al., 2005, p. 567). For example, instead of storing the statement *This person is liberal* combined with a "false" tag, one could simply replace this information by *This person is conservative*. Indeed, in their study, interrupting the encoding of a statement's validity decreased memory for falsity only when the statement was uninformative when being false (e.g., *This person walks barefoot to work*) but not when it was informative when being false (e.g., *This person owns a television*). In a similar vein, Mayo, Schul, and Burnstein (2004) observed memory errors reflecting the loss of "false" tags more often for unipolar statements (i.e., statements lacking a unique opposite) than for bipolar statements that are clearly informative when negated. Taken together, these findings suggest that the Spinozan model probably is restricted to situations in which (1) people lack background knowledge about the presented statements and (2) the statements are uninformative when negated.

The possible role of guessing bias

Although the results of Hasson et al. (2005), Mayo et al. (2004), and Richter et al. (2009) limit the scope of the Spinozan model to some degree, they do not challenge its validity in general. A more fundamental objection concerns ambiguity about the cognitive processes that underlie the feedback interruption effect observed by Gilbert and collaborators. According to the Spinozan model, this effect is caused by different memory representations of true and false propositions. However, there are alternative possible causes. What if guessing biases rather than memory representations

drive the effect? It is indeed conceivable that interruption does not affect memory for falsity per se, but that people tend to guess "true" whenever they are unable to remember the truth value of information. According to the Gricean cooperation principle of conversation, and the "maxim of quality" in particular, "true" should in fact be the best guess for the truth value of propositions encountered in everyday life (Grice, 1989).

Further evidence supporting this interpretation has come from studies addressing the so-called truth effect—that is, the phenomenon that repeatedly presented statements are more likely to be accepted as true than new statements (for a review, see Dechêne, Stahl, Hansen, & Wänke, 2010). A common explanation for the truth effect is that repetition increases processing fluency and that people tend to judge fluently processed statements as true (e.g., Reber & Schwarz, 1999). To test this assumption, Reber and Schwarz manipulated processing fluency using color contrast rather than repetition. Supporting the fluency account, statements presented in clearly discriminable colors received higher truth ratings than statements presented in moderately discriminable colors (see also Unkelbach, 2007). Because all statements were evaluated at their first encounter, differences in memory representations cannot account for the observed truth effect. Therefore, fluency effects are difficult to reconcile with the Spinozan model. However, they could in fact reflect one of the cognitive mechanisms underlying guessing processes. For instance, whenever people have to guess the validity status of a statement, their guesses might be systematically affected by processing fluency or by other metacognitive experiences (cf. Schwarz, Sanna, Skurnik, & Yoon, 2007).

If the results of Gilbert et al. (1990) were really affected by guessing biases, they would not be informative with respect to the memory representations of truth and falsity. Actually, Gilbert and colleagues also considered the possibility of guessing biases when reporting the results of their "Hopi" language experiment. However, because new statements were more often misclassified as false than as true in their experiment, they concluded that there was no "true" guessing bias for unrecognized statements. Although this conclusion seems sound, it is important to note that subsequent replication studies showed a reliable tendency to misremember new items as true (cf. Gilbert et al., 1993), consistent with the guessing hypothesis. Moreover, Gilbert et al. (1990) also denied a "true" guessing bias for recognized statements, because classification times for false statements misidentified as true did not depend on feedback interruption. We doubt that a comparison of decision times suffices to rule out the possibility of guessing biases. Indeed, even Gilbert and colleagues admitted that a guessing bias explanation "cannot be dismissed entirely" (Gilbert et al., 1993, p. 226). Consequently, we will argue that a fair test of the Spinozan and Cartesian models would require measuring memory and guessing processes independently, a goal that can be achieved by means of multinomial processing tree models.

Multinomial processing tree (MPT) models

MPT models are stochastic models for categorical frequency data, as typically obtained in memory experiments (e.g., hits, false alarms, correct rejections, or misses in recognition tests). These models are based on the idea that an observed outcome event (e.g., a hit) does not necessarily reflect a single cognitive process only (e.g., correct item memory). Rather, different underlying processes may contribute to the same event (e.g., memory processes or guessing processes), and each of these processes occurs with a certain probability. Assumptions about the interplay of cognitive processes are represented using simple processing tree diagrams. Such tree models are easily transferable into a set of mathematical model equations that link the probabilities of all possible outcome events to the unknown probabilities of the underlying cognitive processes. Given a set of observed frequency data for the outcome events and an appropriate set of model equations, maximum likelihood methods can be used to assess model fit and to determine probability estimates for the cognitive processes (see Erdfelder et al., 2009, for a recent review). In two experiments, we made use of this approach to disentangle memory processes from guessing processes, and thus to conduct a fair test of the Spinozan model against the Cartesian model. A description of the MPT model we used is provided in the Results section of Experiment 1.

Experiment 1: Encoding of truth and falsity

The Cartesian model presumes that both ostensibly true and ostensibly false information are stored with "true" and "false" tags, respectively. In contrast, according to the Spinozan model, only ostensibly false information is tagged in memory. Hence, the Spinozan coding system should work efficiently in situations in which truth and falsity are mutually exclusive and exhaustive categories. However, confusions should occur when there is a third category of information with unknown validity. In contrast, if the Cartesian model holds, it should always be quite easy to discriminate between true and false information, given that the corresponding tags are stored in memory.

To test these predictions, we conducted a source memory experiment with three different sources. In the study phase, participants read several trivia statements presented by three fictitious persons: Hans, Fritz, and Paul. Following Begg,

Anas, and Farinacci (1992), our experiment involved the following two experimental conditions. Before the study phase, participants in the experimental group (the "precue group") were informed that all statements of Hans were true, that statements of Fritz were true in 50% of the cases, and that all statements of Paul were false. In contrast, participants in the control group (the "postcue group") did not receive this information about Hans, Fritz, and Paul until the test phase. Hence, unlike the participants of the postcue group, participants of the precue group were given the opportunity to encode each statement with a tag indicating its truth value. The test phase was identical for both groups and consisted of an old/new recognition test combined with a source memory test for the three sources, labeled "Hans/true," "Fritz/random," and "Paul/false." Thus, participants in the precue condition could make use of truth value recollection or name recollection to make their source judgments. In contrast, participants in the postcue condition could make use of name recollection only, because validity information had been unavailable in the study phase.

We predicted that if our precue instruction worked as intended (i.e., entailing the encoding of validity information), participants in the precue condition should display better source memory than participants in the postcue condition. This prediction is based on the findings of Begg et al. (1992) that source memory for validity information (available in the precue condition but not in the postcue condition) is superior to source memory for names (available in both conditions). Moreover, if the Cartesian model holds, true statements from Hans and false statements from Paul should be tagged as "true" and "false," respectively, when encoded in the precue condition. In contrast, Fritz's statements of uncertain validity should remain untagged. As a consequence, memory for Hans/true and Paul/false in the precue condition should be equally good, and better than memory for Fritz/random. In contrast, if people store "false" tags only as predicted by the Spinozan model, good source memory in the precue condition should be limited to the false statements of Paul.

Method

Participants A group of 33 participants were randomly assigned to the precue group and the postcue group. The age of the participants ranged from 19 to 29 years (M = 21.94, SD = 2.54). Ten of the participants were male and 23 were female. All participants were University of Mannheim students who received course credit or €4 for participation.

Material We collected 700 true and false trivia statements from different domains (sports, geography, biology, etc.) for pretesting. Each of these statements was evaluated by at least 18 participants on a 7-point truth rating scale, ranging from *definitely false* (1) to *definitely true* (7). Ninety statements with truth ratings between $M = 3.50$ and $M = 4.50$ and standard deviations less than 2 were finally selected as the stimulus materials. Hence, the statements of Experiment 1 were maximally ambiguous with respect to their real truth status (e.g., *Owls are the only birds that can perceive the color blue*). By implication, none of the statements was obviously true or false a priori, so that the effects of previous knowledge observed by Richter et al. (2009) could be ruled out. Moreover, although using trivia statements instead of artificial vocabulary, the informational value of our statements was comparable to the "Hopi" sentences presented by Gilbert et al. (1990). That is, all statements were informative as true statements, but almost no statement was informative when negated.

Selected statements were divided into three stimulus sets so that the mean truth ratings and standard deviations were comparable between sets (Set A: M = 4.01, SD = 1.29; Set B: M = 4.02, SD = 1.29; Set C: M = 4.01, SD = 1.31). Each set consisted of 15 true and 15 false statements. Within each set, ten true statements were assigned to Hans, ten false statements were assigned to Paul, and five true plus five false statements were assigned to Fritz. The statements of Hans, Fritz, and Paul did not differ in their mean truth ratings (Ms = 4.01) and showed comparable standard deviations ($1.28 \leq SD \leq 1.30$).

Design The design comprised two experimental groups, the precue group and the postcue group. Within each group, three stimulus sets were counterbalanced across participants. That is, the statements of sets A and B, B and C, or A and C were presented in the study phase. The statements of the remaining stimulus set (i.e., C, A, or B, respectively) served as distractors during the test phase.

Procedure After signing a consent form and filling out a demographic questionnaire, participants performed the experiment on standard PCs running E-Prime software.

In the first phase of the experiment, they were asked to imagine Hans, Fritz, and Paul playing the quiz game Trivial Pursuit and thus answering different knowledge questions. Participants were informed that all answers would be displayed on the screen and should be memorized along with their respective sources. In addition, participants in the precue group were correctly informed that all of the answers of Hans would be correct (i.e., true), that 50% of Fritz's answers would be correct, whereas 50% would be false, and that all of Paul's answers would be false. Participants in the postcue condition did not receive this information.

After a short practice block, 84 statements and their respective sources were successively presented on the screen. The first 12 and the last 12 statements served as buffer items to prevent primacy and recency effects. The

other 60 statements were randomly drawn from two stimulus sets. Each statement was presented for 6 s in the center of the screen, with the respective source (Hans, Fritz, or Paul) presented above the statement. A 500-ms interstimulus interval preceded the next statement presentation.

The study phase was followed by a 20-min retention interval in which participants performed a nonverbal distractor task. After this interval, the participants of both experimental groups were instructed that all of the statements of Hans had been true, that 50% of the statements of Fritz had been true and 50% false, and that all of the statements of Paul had been false.

In the final phase of the experiment, participants performed a source memory test. A total of 90 statements (60 old and 30 new statements) were randomly presented on the screen. For each statement, participants indicated whether it was old or new. In the case of an "old" judgment, participants also indicated the source of the statement ("Hans/true," "Fritz/random," or "Paul/false").

Results

Mean performance by conditions Statement memory (i.e., proportion of hits minus proportion of false alarms) was comparably good for the precue group ($M = .83$, $SD = .14$) and the postcue group ($M = .82$, $SD = .12$), $t(31) = 0.25$, $p = .80$. Source memory was assessed by means of the conditional source identification measure (CSIM; Murnane & Bayen, 1996). CSIM reflects the proportion of correct source classifications among the correctly recognized target statements. As predicted, mean CSIMs were higher in the precue group than in the postcue group (see Table 1). This finding corroborates our assumption that participants in the precue group indeed based their source memory judgments on validity information rather than on memory for names. The observed group difference in CSIMs was statistically significant, as indicated by a 2 (group: precue, postcue) × 3 (source: Hans/true, Fritz/random, Paul/false) split-plot ANOVA, $F(1, 31) = 36.45$, $p < .001$, $\eta_p^2 = .54$. Moreover, we found no main effect of source, but a significant interaction did emerge between group and source, $F(2, 62) = 12.75$, $p < .001$, $\eta_p^2 = .02$. To analyze the nature of this interaction, we performed separate ANOVAs for the two experimental groups. Whereas CSIMs for the three sources did not differ reliably within the precue group, $F(2, 32) = 2.22$, $p = .13$, they did differ in the postcue group, $F(2, 30) = 17.33$, $p < .001$, $\eta_p^2 = .54$.

Table 1 Mean conditional source identification measures (with standard errors in parentheses) for the precue group and the postcue group of Experiment 1

Group	Source		
	Hans/True	Fritz/Random	Paul/False
Precue	.78 (.04)	.69 (.04)	.73 (.05)
Postcue	.36 (.04)	.53 (.04)	.36 (.05)

Multinomial analyses To disentangle memory processes from guessing processes, the data were additionally analyzed using a slightly modified version of the multinomial three-source MPT model of Riefer, Hu, and Batchelder (1994). In this model, memory processes are represented by the probabilities D (statement recognition) and d (source discrimination), whereas guessing processes are represented by the parameters b, a, and g.

Specifically, the three-source model proposes that an old statement is recognized with probability D. Moreover, if the statement has been recognized, the corresponding source of that statement can be remembered with probability d. If participants do not remember the source of a statement (probability $1 - d$), then they have to guess in order to provide a source judgment. That is, participants will guess with probability $a_{\text{Hans/true}}$ that the statement belongs to the source Hans/true, with probability $a_{\text{Fritz/random}}$ that it belongs to Fritz/random, and with probability $a_{\text{Paul/false}}$ that it belongs to Paul/false. When an old statement is not recognized (probability $1 - D$), in contrast, participants can either correctly guess that the statement is old (with probability b) or wrongly assume that the statement is new (probability $1 - b$). If participants guess "old," then they also have to guess the source of the statement—that is, they guess Hans/true (with probability $g_{\text{Hans/true}}$), Fritz/random (with probability $g_{\text{Fritz/random}}$), or Paul/false (with probability $g_{\text{Paul/false}}$).

Basically, the same process assumptions apply to old test items from all three sources, albeit with the possibility of different source memory parameters. The model for distractor statements, finally, assumes that distractors are detected as new with probability D. If they are not detected as new (probability $1 - D$), the same guessing processes apply as in cases of nonrecognized old statements. For a graphical illustration of the full processing tree model, see Fig. 1.[1] MPT analyses were computed by means of the computer program multiTree (Moshagen, 2010). A likelihood-ratio test showed that the three-source model fit the data well, $G^2(6) = 2.93$, $p = .82$.

[1] Unlike the original model by Riefer et al. (1994), our model is a two-high-threshold model. To keep this model identifiable, we made use of the constraint that the probability D of recognizing an old statement is equal to the probability of detecting a new statement. This is a rather common assumption that has been shown to be appropriate in many applications (see, e.g., Erdfelder et al., 2009; Meiser & Bröder, 2002). Moreover, by means of a model-free χ^2 test (Batchelder & Riefer, 1990), we could show that D does not differ between the three sources of Experiments 1 and 2, $\chi^2(2) \leq 1.33$, $p \geq .57$.

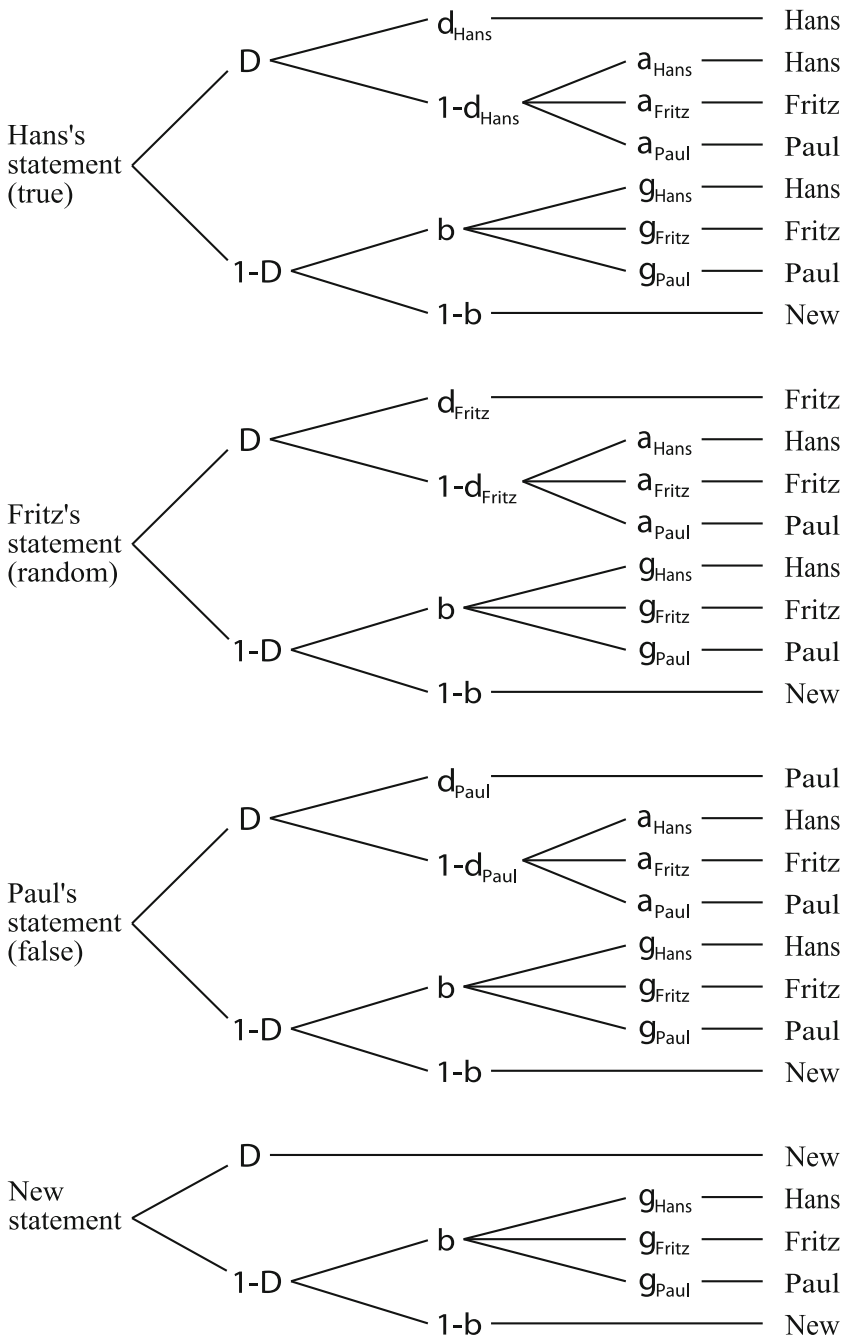

Fig. 1 Structure and parameters of the three-source multinomial model (Riefer et al., 1994) including a two-high-threshold assumption (e.g., Meiser & Bröder, 2002). The model consists of separate processing trees for Hans's statements, Fritz's statements, Paul's statements, and new statements. Each branch of a tree represents a possible sequence of cognitive processes, resulting in one of the following source judgments: "Hans" (true), "Fritz" (random), "Paul" (false), or "new." The parameters in the model reflect transition probabilities: from left to right, D = probability of detecting an old statement as old or a new statement as new; d = probability of detecting the correct source of a statement; b = probability of responding "old" to nondetected statements; a_i = probability of guessing that a statement detected as old belongs to source i; g_i = probability of guessing that a nondetected statement belongs to source i

Memory parameters Statement memory, as measured by D, did not significantly differ between groups, $\Delta G^2(1) = 0.74$, $p = .39$ (see Table 2). In contrast, source memory, as represented by d, clearly differed between and within the two groups. A between-groups comparison revealed a clear source memory advantage of the precue group in contrast to the corresponding parameters of the postcue group, $\Delta G^2(3) = 200.50$, $p < .001$. This finding once more demonstrates that the precue instruction worked as intended. Comparisons of d parameters within the precue group revealed the following results: As predicted by the Cartesian model, source memory did not differ for the true statements of Hans and for the false statements of Paul, $\Delta G^2(1) = 0.30$, $p = .59$. However, source memory for the statements of Fritz, with unknown validity, was considerably lower, $\Delta G^2(2) = 24.97$, $p < .001$, and did not even differ from the overall source memory performance of the postcue group, $\Delta G^2(3) = 3.01$, $p = .39$. Moreover, within the postcue group, we found no significant differences in source memory, $\Delta G^2(2) = 1.74$, $p = .42$.

Guessing parameters Guessing biases for the different sources did not differ between recognized and nonrecognized statements. That is, the a and g parameters could be equated

Table 2 Parameter estimates of the three-source multinomial model (with standard errors in parentheses) for the precue group and the postcue group of Experiment 1

Group	Memory Parameters				Guessing Parameters			
	D	$d_{Hans/true}$	$d_{Fritz/random}$	$d_{Paul/false}$	$a_{Hans/true}$	$a_{Fritz/random}$	$a_{Paul/false}$	b
Precue	.82 (.01)	.73 (.04)	.28 (.11)	.70 (.03)	.24 (.03)	.57 (.05)	.19 (.03)	.10 (.03)
Postcue	.80 (.01)	.11 (.05)	.22 (.06)	.10 (.05)	.30 (.02)	.41 (.03)	.30 (.02)	.14 (.03)

without inducing a significant decrease in model fit, $\Delta G^2(4) = 1.74$, $p = .78$. To keep the model as parsimonious as possible, guessing parameters were therefore estimated under the constraint $a = g$. A comparison of the three a parameters revealed a significant guessing bias in both experimental conditions. Specifically, when participants could not remember the source of a statement, they more often guessed Fritz/random than Hans/true or Paul/false, $\Delta G^2(2) \geq 25.42$, $p < .001$. Guessing probabilities for the sources Hans/true and Paul/false did not differ significantly, $\Delta G^2(2) = 1.56$, $p = .46$. Finally, the b parameters were significantly lower than .50 in both groups, $\Delta G^2(2) = 170.20$, $p < .001$. Thus, in the absence of statement memory, participants tended to guess that a presented statement was new.

Discussion

The goal of Experiment 1 was to investigate the memory representations of truth and falsity by means of a multinomial source memory model. As expected, the precue group displayed a substantial source memory advantage in comparison to the postcue group. This advantage demonstrated the effectiveness of the precue instruction because it implied that participants of the precue group based their source memory judgments on validity information rather than on names. Moreover, within the precue group, source memory did not differ for the true statements of Hans and the false statements of Paul and was much better than source memory for Fritz's statements of uncertain validity. Hence, it can be concluded that source memory is equally good for truth and falsity. This result is in line with the predictions of the Cartesian model. At the same time, it clearly contradicts the Spinozan model. According to the latter view, source memory in the precue group should have been much better for the false statements of Paul than for the true statements of Hans.

One interesting finding of Experiment 1 was the low source memory parameter for the Fritz/random source in the precue group. Why were the participants in this group so much worse at remembering the source of a statement with unknown validity than at remembering the sources of true and false statements? From the Cartesian model, the following explanation seems plausible: Because tags can get lost as a consequence of forgetting, the absence of a memory tag is not diagnostic for the validity status of stored statements. Recognized statements without a tag can be either (a) statements with unknown validity or (b) true or false statements that had initially been stored with a memory tag that had gotten lost over time. Hence, source judgments for untagged statements in the precue group were based on name memory or on guessing. This interpretation is supported by two findings. First, source memory for Fritz/random in the precue group was on the same level as source memory performance in the postcue group. Second, the precue group showed a strong guessing bias toward the Fritz/random response. Apparently, participants in this group held the metacognitive belief that statements whose source could not be remembered were most likely statements with an unknown validity status presented by Fritz. The influence of similar metacognitive inferences on source memory judgments has previously been demonstrated by Meiser, Sattler, and von Hecker (2007).

Experiment 2: Forgetting of truth and falsity

The findings of Experiment 1 show that truth and falsity are encoded equally well, and thus they support the Cartesian model convincingly. However, in contrast to Experiment 1, most real-world situations involve retention intervals of several days or even weeks between the encoding and retrieval of validity information. For instance, in most court cases, jurors and judges are confronted not only with admissible evidence but also with irrelevant, inadmissible evidence, as well as with false media information. Hence, at the end of a trial, a fair conviction can only be attained if decision makers are still able to correctly remember the validity of the information encountered. That this task is not at all easy is demonstrated by several studies revealing a significant influence of inadmissible evidence and media reports on juror verdicts (e.g., Steblay, Besirevic, Fulero, & Jimenez-Lorente, 1999; Steblay, Hosch, Culhane, & McWethy, 2006).

Skurnik, Yoon, Park, and Schwarz (2005) conducted a study in which they directly assessed the influence of

retention interval on memory for validity information. Consistent with our findings in Experiment 1, young participants did not reveal any asymmetries in correct classifications of truth and falsity after a 30-min retention interval. After a three-day interval, however, participants more often misclassified false information as true than vice versa.[2] The authors concluded that, when context memory fades, people more often rely on their metacognitive feelings, such as familiarity or fluency, to infer the truth of a statement. In the multinomial source memory model introduced above, this truth bias should show up as an enhanced "true" guessing bias. Alternatively, however, the observed findings could also reflect a real memory bias. For instance, it is possible that "false" tags are more vulnerable than "true" tags to forgetting.

Thus, even if Descartes is right about the initial representation of truth and falsity, as shown in Experiment 1, differential forgetting of validity information could cause an asymmetry in memory for "true" and "false" tags in the long run. Note that because two types of asymmetry are possible in principle (i.e., faster forgetting of "false" vs. "true" tags), a revised version of the Spinozan model (referring to the retrieval rather than the encoding of "false" tags) could perhaps account for the results observed after longer retention intervals. Clearly, to address this issue, source memory for truth values needs to be assessed for retention intervals of different lengths.

In light of the considerations above, the goal of Experiment 2 was twofold. First, we aimed to replicate our findings from Experiment 1 concerning the encoding of truth and falsity. Second, we wanted to assess whether an asymmetry exists in the forgetting of truth and falsity across ecologically more valid retention intervals up to 1 week. For these reasons, participants again performed the precue task from Experiment 1. However, half of the participants performed the task with a 20-min retention interval (20-min group), and the other half with a 1-week retention interval (1-week group).

Method

Participants A group of 46 participants were randomly assigned to the two experimental conditions. The ages of the participants ranged from 19 to 33 years (M = 21.42, SD = 2.63). Five of the participants were male and 41 were female. All of the participants were University of Mannheim students who received course credit or €5 for participation.

Material The stimulus materials were the same as in Experiment 1.

Design The research design comprised two groups, a 20-min group and a 1-week group. As in Experiment 1, three stimulus sets were counterbalanced across the participants within each group.

Procedure The experiment included two sessions. For the 20-min group, the first session was identical to that for the precue group of Experiment 1. The procedure for the 1-week group was similar. However, unlike the 20-min group, the 1-week group continued working on the nonverbal distractor task after the 20-min retention interval instead of completing a memory test. One week after the first session, participants of both groups returned to the laboratory for a second session. This time, the 1-week group performed the memory test while the 20-min group worked on a distractor task.

Results

Mean performance by conditions Unsurprisingly, statement memory (proportion of hits minus proportion of false alarms) was better after the 20-min retention interval (M = .91, SD = .06) than after the 1-week interval (M = .67, SD = .11). Due to heterogeneity of variances, a Welch test was computed to compare the means. This test indicated that the group difference in recognition performance was statistically significant, $t(33.73) = 9.11$, $p < .001$, Cohen's $d = 2.71$. CSIMs were computed in order to investigate source memory performance. As expected, participants in the 20-min condition had higher CSIMs than did participants in the 1-week condition (see Table 3). This difference was statistically significant, $F(1, 44) = 29.68$, $p < .001$, $\eta_p^2 = .40$, as indicated by a 2 (group: 20-min, 1-week) × 3 (source: Hans/true, Fritz/random, Paul/false) split-plot ANOVA. No other effects were significant.

Multinomial analyses The data were also analyzed with the multinomial three-source model previously used in Experiment 1. Again, we found a good fit to the data, $G^2(6) = 3.77$, $p = .71$.

Table 3 Mean conditional source identification measures (with standard errors in parentheses) for the 20-min group and the 1-week group of Experiment 2

Group	Source		
	Hans/True	Fritz/Random	Paul/False
20-Min.	.73 (.04)	.75 (.04)	.73 (.04)
1-Week	.51 (.04)	.54 (.04)	.45 (.04)

[2] Note that even after 30-min intervals false statements may be misremembered as true (cf. Skurnik, Yoon, & Schwarz, 2007). However, more importantly, the findings of Skurnik et al. (2005) imply that this truth bias increases over time.

Memory parameters A comparison of memory parameters between the groups revealed typical forgetting effects (see Table 4). Specifically, statement memory was significantly lower after the 1-week retention interval than after the 20-min interval, $\Delta G^2(1) = 231.76$, $p < .001$. Moreover, source memory was significantly worse in the 1-week than in the 20-min group, $\Delta G^2(3) = 88.12$, $p < .001$. Within each group, the *d* parameters showed exactly the same pattern previously found for the precue group of Experiment 1. That is, source memory did not differ significantly for the true statements of Hans and the false statements of Paul, $\Delta G^2(1) \leq 0.04$, $p \geq .85$. However, source memory was considerably lower for Fritz's statements, of unknown validity, than for the other two sources, $\Delta G^2(2) \geq 19.21$, $p < .001$.

Guessing parameters Replicating Experiment 1, the *a* and *g* parameters could be equated without a significant decrease in model fit, $\Delta G^2(4) = 2.29$, $p = .68$. Guessing parameters were therefore estimated under the constraint $a = g$. The strong guessing bias toward the Fritz/random source was replicated in Experiment 2: In absence of source memory, participants of both groups more often guessed Fritz/random than either of the two other sources, $\Delta G^2(2) \geq 91.84$, $p < .001$. Guessing probabilities of the sources Hans/true and Paul/false did not differ in the 20-min group, $\Delta G^2(1) = 0.001$, $p = .97$. However, in the 1-week group the tendency to guess Hans/true was significantly higher than the tendency to guess Paul/false, $\Delta G^2(1) = 12.36$, $p < .001$. Finally, and again replicating Experiment 1, the *b* parameters were significantly lower than .50 in both groups, $\Delta G^2(2) = 216.89$, $p < .001$. Thus, in the absence of statement memory, participants tended to guess that a presented statement was new.

Discussion

Experiment 2 produced two major findings. First, within the 20-min condition, source memory did not differ for the true statements of Hans and the false statements of Paul. In contrast, source memory was significantly lower for Fritz's statements of unknown validity than for statements from the other two sources. This nicely replicates the findings for the precue group of Experiment 1, and thus provides further support for the Cartesian model.

Second, as expected, source memory performance was lower in the 1-week group than in the 20-min group. Importantly, however, memory performance for the true statements of Hans and the false statements of Paul differed within neither the 20-min group nor the 1-week group. This finding suggests that "true" and "false" tags are forgotten equally quickly across time.

Consequently, the asymmetries in true–false classifications previously observed by Skurnik et al. (2005) are in fact due to guessing biases rather than memory biases. Indeed, an inspection of guessing tendencies showed that participants in the 1-week condition were more likely to attribute statements to the Hans/true source than to the Paul/false source. Interestingly, this guessing bias did not depend on actual statement recognition, as was indicated by the fact that the *a* and *g* parameters did not differ. In contrast, all statements classified as "old" (i.e., even the new ones) were more likely to be attributed to the Hans/true than to the Paul/false source. One possible explanation is that "old" judgments and "true" judgments are affected by the same metacognitive feelings (e.g., familiarity and processing fluency). For example, preexperimental familiarity could not only explain why people judge new statements as old, but also why they tend to judge these statements as true. Indeed, several studies have shown that "statements that are judged to be repeated are rated as truer than statements judged to be new, regardless of the actual status of the statements" (Bacon, 1979, p. 241; see also Hawkins & Hoch, 1992; Law, 1998).

General discussion

The main purpose of our experiments was to investigate the memory representations of truth and falsity implied by the Spinozan and Cartesian models. Both models predict that apparently false statements will be stored in memory with "false" tags. However, the models make different predictions concerning the memory representations of apparently true statements. Whereas the Cartesian model assumes that those statements will be stored along with "true" tags, the

Table 4 Parameter estimates of the three-source multinomial model (with standard errors in parentheses) for the 20-min group and the 1-week group of Experiment 2

Group	Memory Parameters				Guessing Parameters			
	D	$d_{\text{Hans/true}}$	$d_{\text{Fritz/random}}$	$d_{\text{Paul/false}}$	$a_{\text{Hans/true}}$	$a_{\text{Fritz/random}}$	$a_{\text{Paul/false}}$	*b*
20-Min.	.91 (.01)	.68 (.03)	.35 (.09)	.68 (.03)	.18 (.02)	.63 (.04)	.19 (.02)	.08 (.03)
1-Week	.63 (.02)	.36 (.05)	.00 (.10)	.38 (.04)	.28 (.02)	.55 (.03)	.18 (.02)	.21 (.02)

Spinozan model proposes that they will remain untagged. By means of a multinomial source memory model, we showed that our data support the Cartesian model and contradict the Spinozan model, both for short (Exp. 1) and for long (Exp. 2) retention intervals.

Obviously, our results are at odds with the conclusions of Gilbert et al. (1990), whose "Hopi" language experiment provided evidence compatible with the Spinozan model. However, unlike those of our experiments, the findings of Gilbert and colleagues were solely based on the proportions of correct source classifications (SIM) as the dependent variable. One severe problem of both SIM and the related CSIM is that source memory performance and guessing tendencies are confounded in a single score (Bayen, Murnane, & Erdfelder, 1996; Bröder & Meiser, 2007; Murnane & Bayen, 1996; Vogt & Bröder, 2007). By implication, it remains unclear whether the asymmetry in correct classifications of true and false statements observed in the "Hopi" language experiment can really be attributed to source memory differences for truth and falsity, as suggested by Gilbert and colleagues. Given our multinomial modeling results, we doubt this interpretation, even more so because a comparison of CSIM and multinomial modeling results corroborates our argument that conventional source memory measures are heavily influenced by guessing biases. For example, our findings in Experiment 2 suggest that people tend to guess "true" more often than "false" when context memory fades. It therefore appears likely that the truth guessing bias also increases when context memory is impaired by feedback interruption. This supports our view that the feedback interruption effect in Gilbert et al.'s (1990) "Hopi" language experiment was caused by a "true" guessing bias in the first place, rather than by different memory representations of truth and falsity.

Although our results provide strong support for the Cartesian model, we think they should only be seen as a first step toward isolating the memory representations of truth and falsity using a model-based approach. In addition to examining the forgetting of "true" and "false" tags, which we accomplished in Experiment 2, other open questions will need to be addressed in subsequent studies.

First, in most real-world situations, the amount of true information encountered clearly exceeds the amount of false information. In our experiments as well as in the "Hopi" language experiment, in contrast, the base rates of true and false information did not differ. However, the proportion of true information could be a crucial factor that affects the memory representations of truth and falsity. For example, whenever false information is rare, people might switch from the Cartesian to the Spinozan tagging system. This would minimize cognitive effort, because the majority of information would not require tags. Moreover, different base rates of truth and falsity could trigger feelings of trust or distrust, which in turn might affect information encoding (as was previously demonstrated by Schul, Mayo, & Burnstein, 2004). Similarly, information encoding could also be influenced by other context factors. For instance, it could make a difference whether validity information is indicated by source credibility (as in our experiments) or whether it is communicated using more neutral forms of feedback (as in the "Hopi" language experiment). Thus, it is clearly important to examine the memory representations of truth and falsity under different context conditions. In doing so, potential context effects on memory should be thoroughly disentangled from context effects on guessing.

Second, as demonstrated by the studies of Richter et al. (2009) and Hasson et al. (2005), both relevant background knowledge and the informational value of statements affect memory judgments for truth and falsity. These findings suggest that there may be different memory representations of truth and falsity, depending on what type of statement is processed. Unfortunately, however, it is unclear whether biased guessing also contributes to such stimulus effects. For example, in the Hasson et al. study, statements that were informative when false mainly described familiar characteristics (e.g., *this person owns a television*), whereas statements that were uninformative when false often described unusual habits or characteristics (e.g., *this person walks barefoot to work*). These differences in familiarity or novelty may have caused differences in statement processing (e.g., deeper encoding of unexpected, and thus distinct, statements) or guessing (e.g., a "true" guessing bias for familiar statements) that in turn may have influenced the results. At this point, of course, we do not want to imply that all findings can be attributed to biased guessing. Rather, we argue that subsequent studies should pay special attention to effects of the materials and the context conditions both on the memory representations of truth and falsity and on guessing processes. This will be best achieved using modeling techniques such as MPT models that systematically disentangle the contributions of memory and guessing on cognitive judgments.

Third, in its present form the Cartesian model does not specify whether or not it is possible to change the tag of a statement's memory representation retrospectively. What happens, for example, when information has been accepted as true in the first place but is later uncovered as being false? Does retraction of a statement induce a retagging process, or does it result in the construction of a new, coexisting memory representation (cf. Ecker, Lewandowsky, Swire, & Chang, 2011)? Questions like these demonstrate the need for more elaborated theoretical models and for additional empirical studies. Importantly, however, because of the conclusive evidence presented here, these additional steps should not build upon the Spinozan model, which is inconsistent with our data. Overall, the Cartesian model appears to provide a more appropriate framework for an elaborated theory of truth value representations in human memory.

Author note The reported experiments were part of the first author's German dissertation "Die Wahrheitsillusion" (ISBN: 978-3-89574-709-0).

References

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 241–252.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97,* 548–564. doi:10.1037/0033-295X.97.4.548

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 22,* 197–215. doi:10.1037/0278-7393.22.1.197

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology. General, 121,* 446–458.

Bennett, J. F. (1984). *A study of Spinoza's ethics*. Indianapolis: Hackett.

Bröder, A., & Meiser, T. (2007). Measuring source memory. *Journal of Psychology, 215,* 52–60. doi:10.1027/0044-3409.215.1.52

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

Descartes, R. (1984). Fourth meditation. In J. Cottingham, R. Stoothoff, & D. Murdoch (Eds.), *The philosophical writings of Descartes* (Vol. 2, pp. 37–43). Cambridge: Cambridge University Press (Original work published 1641).

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review, 18,* 570–578. doi:10.3758/s13423-011-0065-1

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology, 217,* 108–124.

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

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,* 601–613.

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,* 221–233.

Grice, H. P. (1989). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 22–40). Cambridge: Harvard University Press.

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

Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research, 19,* 212–225.

Koslow, S., & Beltramini, R. F. (2002). Consumer skepticism and the "waiting room of the mind": Are consumers more likely to believe advertising claims if they are merely comprehended? *Advances in Consumer Research, 29,* 473–479.

Law, S. (1998). Do we believe what we remember or, do we remember what we believe? *Advances in Consumer Research, 25,* 221–225.

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

Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28,* 116–137. doi:10.1037/0278-7393.28.1.116

Meiser, T., Sattler, C., & von Hecker, U. (2007). Metacognitive inferences in source memory judgements: The role of perceived differences in item recognition. *Quarterly Journal of Experimental Psychology, 60,* 1015–1040.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42,* 42–54. doi:10.3758/BRM.42.1.42

Murnane, K., & Bayen, U. J. (1996). An evaluation of empirical measures of source identification. *Memory & Cognition, 24,* 417–428. doi:10.3758/BF03200931

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8,* 338–342.

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

Riefer, D. M., Hu, X., & Batchelder, W. H. (1994). Response strategies in source monitoring. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 20,* 680–693. doi:10.1037/0278-7393.20.3.680

Schul, Y., Mayo, R., & Burnstein, E. (2004). Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *Journal of Personality and Social Psychology, 86,* 668–679.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology, 39,* 127–161.

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research, 31,* 713–724.

Skurnik, I., Yoon, C., & Schwarz, N. (2007). *Education about flu can reduce intentions to get a vaccination*. Unpublished manuscript.

Spinoza, B. (2006). *The ethics*. Middlesex: Echo Library (Original work published 1677).

Steblay, N. M., Besirevic, J., Fulero, S. M., & Jimenez-Lorente, B. (1999). The effects of pretrial publicity on juror verdicts: A meta-analytic review. *Law and Human Behavior, 23,* 219–235.

Steblay, N., Hosch, H. M., Culhane, S. E., & McWethy, A. (2006). The impact on juror verdicts of judicial instruction to disregard inadmissible evidence: A meta-analysis. *Law and Human Behavior, 30,* 469–492.

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33,* 219–230.

Vogt, V., & Bröder, A. (2007). Independent retrieval of source dimensions: An extension of results by Starns and Hicks (2005) and a comment on the ACSIM measure. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33,* 443–450.