Memory & Cognition (2019) 47:1386–1400
https://doi.org/10.3758/s13421-019-00940-6

More evidence against the Spinozan model: Cognitive load diminishes memory for “true” feedback

Lena Nadarevic[1] · Edgar Erdfelder[1]

Published online: 18 June 2019

Abstract

We tested two competing models on the memory representation of truth-value information: the Spinozan model and the Cartesian model. Both models assume that truth-value information is represented with memory “tags,” but the models differ in their coding scheme. According to the Cartesian model, true information is stored with a “true” tag, and false information is stored with a “false” tag. In contrast, the Spinozan model proposes that only false information receives “false” tags. All other (i.e., untagged) information is considered as true by default. Hence, in case of cognitive load during feedback encoding, the latter model predicts a load effect on memory for “false” feedback, but not on memory for “true” feedback. To test this prediction, participants studied trivia statements (Experiment 1) or nonsense statements that allegedly represented foreign-language translations (Experiment 2). After each statement, participants received feedback on the (alleged) truth value of the statement. Importantly, half of the participants experienced cognitive load during feedback processing. For the trivia statements of Experiment 1, we observed a load effect on memory for both “false” *and* “true” feedback. In contrast, for the nonsense statements of Experiment 2, we found a load effect on memory for “true” feedback *only*. Both findings clearly contradict the Spinozan model. However, our results are also only partially in line with the predictions of the Cartesian model. For this reason, we suggest a more flexible model that allows for an optional and context-dependent encoding of “true” tags and “false” tags.

Keywords Truth bias · Spinoza · Descartes · Feedback memory · Multinomial model

Information processing is an essential part of everyday life. However, not every piece of information that people encounter is reliable (e.g., gossip, social media postings, reports of dubious media). Especially in today’s so-called postfactual age, people often encounter misleading or even completely false information (aka *fake news*). In order to protect people from fake news, the social media platform Facebook launched a fact-checking tool in 2017 that tags disputed postings with a warning label (Hunt, 2017). By contrast, however, correct information is not tagged with a specific label. This tagging scheme used by Facebook corresponds to truth-value coding in accordance with what Gilbert, Krull, and Malone (1990) called the *Spinozan model* of the mind. In contrast, it is incompatible with what they termed the *Cartesian model* of the mind.

Spinozan versus Cartesian truth-value tagging

The Spinozan model builds on the assumption of the philosopher Baruch Spinoza that “will and understanding are one and the same” (Spinoza, 1677/2006, p. 52), implying that newly acquired information is believed initially (Bennett, 1984). Based on this view, Gilbert et al. (1990) set up the following predictions for a Spinozan model of the mind: First, new information is primarily believed and represented in memory. Second, if a person has sufficient cognitive capacity to evaluate the information in a later processing step or encounters evidence on its factual truth value, its memory representation may be subsequently tagged as “false.” Accordingly, the Spinozan model distinguishes between untagged and tagged statement representations. The former ones are considered “true” and the latter ones “false.” Because only false statements need to be tagged, the Spinozan model represents truth-value information in an economic way. However, this tagging scheme is prone to errors. For instance, false information may not be tagged as such if cognitive resources for the second processing step are depleted. As a consequence, false information may later be remembered as “true.”

✉ Lena Nadarevic
nadarevic@psychologie.uni-mannheim.de

[1] Department of Psychology, School of Social Sciences, University of Mannheim, D-68131 Mannheim, Germany

A second, contrasting model proposed by Gilbert et al. (1990) refers to the claim of the philosopher René Descartes that people do not automatically believe new information in the first place: "We have free will, enabling us to withhold our assent in doubtful matters and hence avoid error" (Descartes, 1644/1985, p. 194). Descartes' idea led Gilbert and collaborators to set up the following predictions for a Cartesian model of the mind: First, new information is initially represented in memory without any reference to its truth value. Second, if a person has sufficient capacity to evaluate the information or encounters evidence on its factual truth value in a later processing step, it may subsequently be tagged as either "true" or "false." Accordingly, the Cartesian model distinguishes between three different truth-value representations: untagged statements, statements tagged as "true," and statements tagged as "false." It follows that the Cartesian model represents truth-value information in a less economic way than the Spinozan model does. Obviously, this representation is less error prone, because untagged false information is not automatically assumed to be true.[1]

Empirical tests of the Spinozan and the Cartesian model

To test the two models against each other, Gilbert and colleagues (Gilbert et al., 1990; Gilbert, Tafarodi, & Malone, 1993) conducted several experiments in which they induced cognitive load while participants received truth-value feedback on previously presented propositions. Gilbert et al.'s rationale for choosing this paradigm was the following: Because the Spinozan model does not include "true" tags, the load manipulation should only interfere with the encoding of "false" tags. That is, the Spinozan model predicts that cognitive load at feedback encoding selectively impairs memory for "false" feedback. In contrast, if the Cartesian model is correct, and a person stores "false" and "true" tags, then cognitive load should instead interfere with the encoding of both kinds of tags. Hence, the Cartesian model predicts that cognitive load during feedback processing reduces memory for both "false" feedback *and* "true" feedback.

In their first experiment—the so-called Hopi language study—Gilbert et al. (1990) presented their participants with nonsense statements such as "*a monishna is a star*," each of which allegedly consisted of a word of the Hopi Indian language (e.g., *monishna*) and its English translation (e.g., *star*). For two thirds of the statements, participants also received an alleged truth-value feedback ("true" vs. "false") immediately after statement presentation. In some of the trials, feedback processing was distracted by an unrelated stimulus–response task to investigate cognitive load effects on feedback encoding. The results of the Hopi experiment and later studies of Gilbert and colleagues (Gilbert et al., 1990; Gilbert et al., 1993) showed that cognitive load during feedback encoding reduced the proportion of correct "false" attributions in a memory test, but not the proportion of correct "true" attributions. Apparently, the cognitive load manipulation selectively impaired memory for "false" feedback. This pattern of results was expected under the Spinozan model and appears to be inconsistent with the Cartesian model. Gilbert et al. (1990) thus rejected the latter model: "Rene Descartes was right about so many things that he surely deserved to be wrong about something: How people come to believe certain ideas and disbelieve others may be the something about which he was mistaken" (p. 601).

However, the authors' conclusion might have been premature. Subsequent studies investigated the robustness of Gilbert et al.'s (1990) findings with different statement types. Hasson, Simmons, and Todorov (2005) used short person descriptions that differed in their informational value when being false. The authors only replicated Gilbert et al.'s findings for statements that were uninformative when being false (e.g., *this person walks barefoot to work*), but not for statements that were informative when being false (e.g., *this person is liberal*). The authors argued that in case of "false" feedback, the latter statements are represented as affirmative inferences (e.g., *this person is conservative*) instead of propositions attached with "false" tags (cf. Glenberg, Robertson, Jansen, & Johnson-Glenberg, 1999, on the context-dependent informational value of negated propositions). Furthermore, Richter, Schroeder, and Wöhrmann (2009) only found evidence for the Spinozan model when presenting statements for which participants had no or weak background knowledge (e.g., *toothpaste contains sulfur*), but not when using statements for which participants had strong background knowledge (e.g., *soft soap is edible*). Richter and colleagues concluded that strong background knowledge enables a fast validation process that prevents people from automatically accepting all information they encounter as being true. In line with this reasoning, recent studies found evidence for an automatic, knowledge-based validation process (Isberner & Richter, 2013; Piest, Isberner, & Richter, 2018; but see Wiswede, Koranyi, Müller, Langner, & Rothermund, 2013). Taken together, these findings limit the scope of the Spinozan model to statements that are (a) uninformative if they are false and (b) not linked to strong background knowledge.

What is more, Nadarevic and Erdfelder (2013) criticized previous studies for assessing memory for truth-value feedback by comparing the proportion of correct "true" and "false" feedback attributions. Because this proxy measure of memory performance is not process pure (because it

[1] Note that although Gilbert et al.'s Spinozan model and Cartesian model build on ideas of Spinoza and Descartes, they go beyond these ideas by proposing specific memory "tagging" schemes of truth values not explicitly considered by the philosophers.

confounds feedback memory with item memory and possible guessing influences), it can lead to biased results and thus to false conclusions, as is outlined in the relevant source-memory literature (Bayen, Murnane, & Erdfelder, 1996; Bröder & Meiser, 2007; Murnane & Bayen, 1996; Riefer, Hu, & Batchelder, 1994; Vogt & Bröder, 2007). For this reason, Nadarevic and Erdfelder argued that the higher proportion of correct "true" attributions could possibly reflect a "true" guessing bias in case of memory uncertainty rather than actual differences in memory for "true" versus "false" feedback. According to the Gricean maxim of quality (Grice, 1989), the best guess for recently acquired information of uncertain truth status is "true," at least if people trust in the cooperation principle in communication. Empirical support for the guessing account comes from a study by Street and Kingstone (2016). The authors found that cognitive load during feedback encoding increased the proportion of "true" responses at test only when response options were limited to "true" and "false," but not when participants had the additional response option "unsure." The "unsure" option possibly absorbed cases of truth status uncertainty, thus eliminating the bias to guess "true."

Based on their critique of previous studies, Nadarevic and Erdfelder (2013) used a multinomial source monitoring model to disentangle memory for truth-value information from guessing processes (see Erdfelder et al., 2009, for a review on multinomial models and Appendix 1 for a brief introduction). Participants studied trivia statements (e.g., *Manama is the capital of Bahrain*) that were presented by three sources that differed in credibility: "Hans" actually presented only true statements, "Fritz" presented an equal number of true and false statements, and "Paul" actually presented only false statements. Participants in the so-called precue group were informed about the credibility of the three sources before the study phase. Compared with a control group, who received the same three names without the associated credibility information in the study phase, the precue group showed improved source memory for both certainly true statements (i.e., Hans's statements) and certainly false statements (i.e., Paul's statements) in a later source memory test. Obviously, participants in the precue group had encoded the truth-value of these statements (i.e., true vs. false) rather than the names of the sources that were kept constant across conditions. Moreover, memory for truth and falsity did not differ, whereas memory for uncertainty (i.e., source memory for Fritz's statements) was much lower. Taken together, the findings support the prediction of the Cartesian model that statements encoded as true are stored with "true" tags, whereas statements encoded as false are stored with "false" tags.

However, one critical limitation of the study of Nadarevic and Erdfelder (2013) is the lack of a direct test of the processing hypotheses underlying the Spinozan model and the Cartesian model, respectively. More precisely, Nadarevic and Erdfelder did not investigate the effect of cognitive load on memory for truth-value feedback, a crucial element in Gilbert et al.'s (Gilbert et al., 1990; Gilbert et al., 1993) theory. According to Gilbert and collaborators, cognitive load should selectively impair memory for "false" feedback, but not for "true" feedback, if the Spinozan model holds. In contrast, it should impair memory for both types of feedback if the Cartesian model holds.

To conduct a direct test of these competing hypotheses, we conceptually replicated Gilbert et al.'s (1990, Experiment 1) procedure in our current research. More precisely, we presented participants with different statements, each of which was followed by a truth-value feedback as in Gilbert et al.'s research. The feedback was processed either while performing a secondary task (interruption group) or without any distraction (control group). Importantly, however, we analyzed the data of a subsequent memory test in two different ways: In line with Gilbert et al., we analyzed the proportion of correct feedback attributions at first. In addition, however, we analyzed the data with a multinomial source monitoring model. The latter approach has substantial benefits compared with the former because it provides feedback memory estimates unconfounded by statement memory and guessing processes (Bayen et al., 1996; Bröder & Meiser, 2007; Vogt & Bröder, 2007). Moreover, to enhance the generalizability of our findings, we tested the abovementioned predictions of the Spinozan model and the Cartesian model with different statement types: Experiment 1 used the trivia statements of Nadarevic and Erdfelder (2013), whereas Experiment 2 used the Hopi language statements of Gilbert et al. (1990).

Experiment 1

Participants studied trivia statements and received truth-value feedback for each statement ("true," "false," or "uncertain"). Similar to Gilbert et al. (1990), we induced cognitive load at feedback encoding by means of a secondary task. Specifically, participants in the *interruption group* had to perform an unrelated visual discrimination task during feedback encoding. In contrast, in the *control group* feedback encoding was not interrupted by another task. A subsequent memory test assessed (a) statement recognition ("old"/"new") and (b) feedback recognition ("true"/ "false"/ "uncertain") for all statements judged as "old." We used this two-step recognition procedure because remembering the truth-value feedback of a certain statement requires remembering the statement in the first place.

In contrast to previous studies (Gilbert et al., 1990; Gilbert et al., 1993; Hasson et al., 2005; Pantazi, Kissine, & Klein, 2018; Richter et al., 2009), we relied not only on proxy measures of feedback memory but additionally employed a more fine-graded multinomial processing tree model (MPT model) to measure feedback memory unconfounded by statement memory and guessing. The MPT model we used is an adaptation of Riefer et al.'s (1994) source monitoring model for three sources of test statements previously presented in the encoding phase (true, false, and uncertain statements) and new statements presented only in the test phase. As illustrated in Fig. 1, the model assumes that participants recognize a statement from the study phase with probability *D*. If participants recognize a statement, they can also remember the corresponding feedback with probability *d*. However, if they cannot remember the feedback (probability 1 − *d*), they will have to guess the feedback in order to provide a response. That is, participants will guess with probability a_{true} that the statement was presented as "true," with probability a_{false} that it was presented as "false," and with probability $a_{\text{uncert.}}$ that it was presented as "uncertain" in the study phase. If participants do not recognize a statement from the study phase (probability 1 − *D*), participants will either correctly guess that the statement is old (probability *b*), or they will incorrectly assume that the

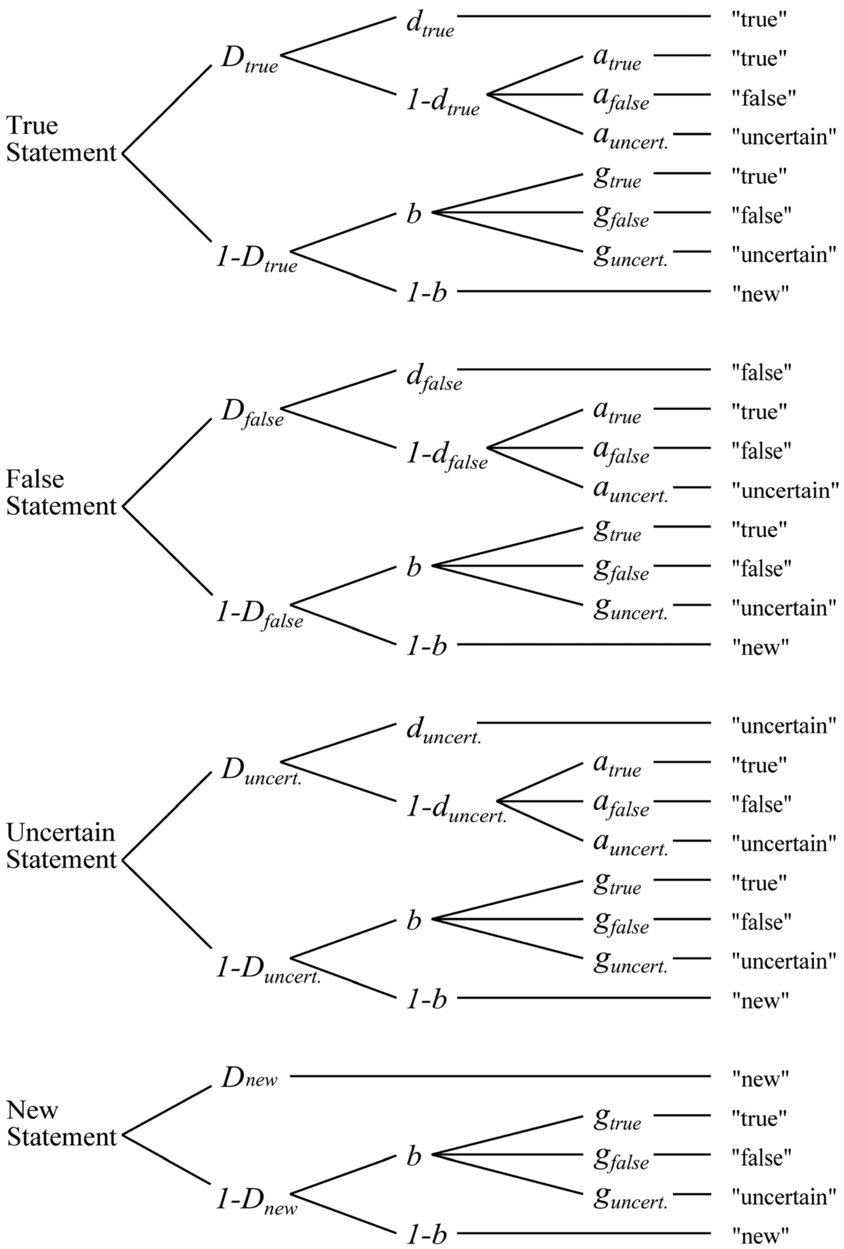

Fig. 1 Structure and parameters of the two-high-threshold variant of the three-sources MPT model of Riefer et al. (1994). The model consists of separate processing trees for statements with true feedback, false feedback, and uncertain feedback, as well as for new statements. Each branch of a tree represents a possible sequence of cognitive processes resulting in a "true," "false," "uncertain," or "new" response. Parameters in the model reflect transition probabilities from left to right: *D* = probability of statement recognition or lure detection, respectively; *d* = probability of feedback recognition; *b* = probability of guessing "old" in case of recognition uncertainty; a_i = probability of guessing feedback *i* for recognized statements; g_i = probability of guessing feedback *i* for unrecognized statements

statement is new (probability $1 - b$). In case of an "old" guess, participants will also have to guess the feedback presented along with the statement. That is, they will guess with probability g_{true} that the statement was presented as "true," with probability g_{false} that it was presented as "false," and with probability $g_{\text{uncert.}}$ that it was presented as "uncertain" in the study phase.

In sum, the MPT approach we employed allows us to disentangle, to estimate, and to compare the following probabilities: Probability of statement recognition (*D*-parameter), probability of feedback recognition (*d*-parameter), probability of guessing "old" in case of recognition uncertainty (*b*-parameter), and, finally, probability of guessing a certain truth-value feedback for recognized statements (*a*-parameters), as well as for unrecognized statements (*g*-parameters). A more detailed introduction to MPT modeling in general and the applied MPT model in particular is provided in Appendix 1 (see also Klauer & Wegener, 1998, Appendix 1, for an easy-to-understand introduction to MPT modeling).

Method

A minimum sample size of $N = 60$ was determined a priori. A sensitivity analysis with G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) revealed that, given this sample size, a conventional α-level of .05, and a target-power of $1 - \beta = .99$, the goodness-of-fit test for the described MPT model will be powerful enough to detect even quite small deviations from the model according to Cohen's (1988) effect size conventions (i.e., $w = .06$). All data exclusions, all manipulations, and all measures employed in the study are reported below.

Participants Sixty-six University of Mannheim students participated for course credit or voluntarily. One participant of the interruption group did not respond to the visual discrimination task and was therefore excluded from all analyses. Thus, the remaining sample consisted of 65 participants (50 female, 15 male) with a mean age of $M = 22.0$ ($SD = 4.5$) years.

Material We used the 90 trivia statements of Nadarevic and Erdfelder (2013) as statements for the study phase and test phase, respectively. These statements were divided in three sets, each containing 15 true statements (e.g., *Manama is the capital of Bahrain*) and 15 false statements (e.g., *Robbie Williams's middle name is Maximilian*)—that is, 30 statements in total. Within each set, 10 true statements were assigned to the feedback "true," 10 false statements to the feedback "false," and the remaining 10 statements (five true ones and five false ones) to the feedback "uncertain." Statements' mean pretested validity ratings ranged between 3.5 and 4.5 ($M = 4.01$) on a 7-point scale and were very similar between sets as well as between "true," "false," and "uncertain" statements within each set. Another set of 24 trivia statements served as stimulus material for the practice and buffer trials. Note also that the "true" and "false" feedback provided was always in agreement with the actual truth value of the respective statement.

Procedure After signing a consent form, the computer instructed participants to imagine themselves as prospective trivia game show candidates. Participants were told that in order to practice for the show, they would have to memorize the truth status of different trivia statements. Participants in the interruption group were also informed that while studying they would oftentimes be interrupted by their mother (as indicated by a picture of a woman) or their little brother (as indicated by a picture of a boy) and had to respond to these interruptions.

Each trial of the study phase started with a fixation cross appearing for 1 second in the center of the screen. Next, a trivia statement was presented for 3 seconds. Subsequently, feedback on the truth status of the statement ("true" vs. "false" vs. "uncertain") replaced the statement and was displayed for 3 seconds. In the interruption group, 750 ms after feedback onset (identical to the delay of Gilbert et al., 1990, Experiment 1), a picture of either a woman or a boy appeared on the left-hand or right-hand side of the feedback. In each trial, it was randomly determined which of the two pictures appeared (woman vs. boy) and at which position the picture appeared (left vs. right). Participants were instructed to respond as fast as possible to each picture by pressing a left key ("d" for the woman) or right key ("k" for the boy). Each picture was displayed until participants responded or until the end of feedback presentation. In the control group, feedback processing was not interrupted by an additional task.

The study phase, which started after four practice trials, comprised 80 trials in total. The first 10 and last 10 trials served as buffer trials. In the 60 middle trials, the statements of two stimulus sets were presented in random order. Following the study phase, participants solved Sudoku puzzles as a nonverbal distractor task for 20 minutes. In the final memory test, 90 statements (60 old ones—i.e., 20 per feedback type—and 30 new ones) were presented in random order. For each statement, participants had to indicate whether it was old or new. In case of an "old" response, they also had to indicate the feedback for the statement ("true," "false," or "uncertain"). Finally, a questionnaire asked participants whether they had used an encoding strategy in the study phase, and, if so, to describe their strategy.

Design Feedback type ("true," "false," "uncertain") was manipulated within participants, whereas cognitive load at feedback encoding was manipulated between participants.

Participants were randomly assigned to the interruption group ($n_1 = 32$) and the control group ($n_2 = 33$). Moreover, statement sets and distractor set at test were counterbalanced across participants.

Results

All statistical tests are based on an α-level of .05. If the sphericity assumption for within-subjects effects was violated (as indicated by Mauchly's test), Greenhouse–Geisser-corrected degrees of freedom were used.

Visual discrimination task performance Participants in the interruption group attended to the visual discrimination task as indicated by a very high performance in this task (proportion of correct responses: *M* > .99, *SD* = .01). Moreover, performance in the discrimination task was unaffected by the feedback type ("true," "false," or "uncertain") processed simultaneously, *F* < 1.

Memory test performance by condition First, we assessed statement memory by calculating the proportion of hits (old statements correctly indicated as "old" in the recognition test) corrected for the proportion of false alarms (new statements mistakenly indicated as "old")—that is, proportion of hits minus proportion of false alarms. Because the discrimination task in the interruption condition was always performed simultaneously with feedback processing—but not with statement processing—we did not expect any differences in statement memory between the interruption group and the control group. Indeed, statement memory was very high in both groups and did not significantly differ between the two (interruption group: *M* = .89, *SD* = .11; control group: *M* = .92, *SD* = .09), *t*(63) = 0.87, *p* = .39, *d* = 0.22.

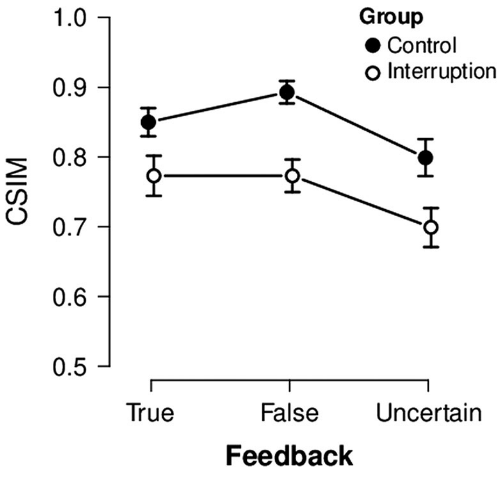

Fig. 2 Mean CSIMs as a function of feedback type and group in Experiment 1. Error bars represent standard errors

Feedback memory was assessed by the proportion of correct feedback attributions among the correctly recognized target statements, which corresponds to the conditional source identification measure (CSIM; Murnane & Bayen, 1996). We used this proxy because it is less confounded by statement memory than the simpler source identification measure (SIM), which in our case reflects the proportion of correct feedback attributions among all target statements (for reviews, see Bröder & Meiser, 2007; Murnane & Bayen, 1996). The descriptive CSIM results are displayed in Fig. 2 (for exact CSIMs, see Table 1). A 3 (feedback type: true vs. false vs. uncertain) × 2 (group: interruption vs. control) ANOVA on mean CSIMs revealed a main effect of group, *F*(1, 63) = 16.37, *p* < .001, η_p^2 = .21. This effect reflects the expected cognitive load effect of the visual discrimination task on feedback memory as indicated by lower CSIMs in the interruption group as compared with the control group. Moreover, CSIMs differed between the three feedback types, *F*(1.83, 115.01) = 8.96, *p* < .001, η_p^2 = .12. Post hoc tests with Bonferroni-adjusted *p* values showed that this main effect was due to lower CSIMs for "uncertain" feedback as compared with "true" feedback, *t*(64) = 2.78, $p_.$ = .021, *d* = 0.35, and "false" feedback, *t*(64) = 3.90, *p* < .001, *d* = 0.48. CSIMs for "true" and "false" feedback did not differ significantly, *t*(64) = 1.25, *p* = .645, *d* = 0.16. Importantly, there was no Group × Feedback Type interaction, *F* < 1. This finding shows that cognitive load at feedback encoding reduced the proportion of correct "false" *and* correct "true" attributions at test. The following planned comparisons supported this conclusion: CSIMs for "false" feedback were significantly lower in the interruption group compared with the control group, *t*(63) = 4.28, *p* < .001 (one-tailed), *d* = 1.06. Likewise, CSIMs for "true" feedback were also significantly lower in the interruption group compared with the control group, *t*(63) = 2.18, *p* = .017 (one-tailed), *d* = 0.54. Notably, this data pattern contradicts the results of Gilbert et al. (1990, Experiment 1), who had found a cognitive load effect on "false" feedback attributions, but not on "true" feedback attributions.

Table 1 Mean CSIMs (with standard errors) for the control group and the interruption group of Experiment 1

	Feedback		
	True	False	Uncertain
Control group	.85 (.02)	.89 (.02)	.80 (.03)
Interruption group	.77 (.03)	.77 (.02)	.70 (.03)

Multinomial analyses In order to disentangle statement memory, feedback memory, and guessing processes, data were also analyzed with a two-high-threshold variant of the MPT model of Riefer et al. (1994), explained above (see Fig. 1). For the current experiment, the model was specified as follows: To assess group differences in statement memory (*D*-parameter), *D* was estimated separately for the interruption group and the control group. Moreover, *D* was also estimated separately for statements with "true," "false," and "uncertain" feedback. To make sure that the model is identifiable, we implemented a well-established parameter restriction on the D_{new}-parameter that is in line with the mirror effect—that is, the tendency for correct rejections to increase (or decrease, respectively) symmetrically with hit rates (Glanzer, Adams, Iverson, & Kim, 1993). More precisely, we constrained the probability of lure detection (D_{new}-parameter) to be equal with the probability of recognizing statements with "uncertain" feedback (D_{uncert}-parameter) within each group (see Bell, Buchner, & Musch, 2010, for an equivalent restriction).[2] Memory for truth-value feedback (*d*-parameter)—the crucial parameter for testing the Spinozan against the Cartesian account—was also estimated separately for the interruption group and the control group as well as for the different feedback types. Moreover, all guessing parameters (i.e., the *b*-parameter, *a*-parameters, and *g*-parameters) were estimated separately for the two experimental groups. MPT analyses were computed with multiTree (Moshagen, 2010). A likelihood-ratio test confirmed that there was no significant model misfit, $G^2(2) = 5.75$, $p = .057$. In other words, the observed response frequencies were in agreement with the model's predictions. Parameter estimates of the model are summarized in Table 2.

Statement memory. In the control group, statement memory (i.e., *D*-parameters) was affected by feedback type, $\Delta G^2(2) = 6.55$, $p = .038$. Pairwise comparisons of the *D*-parameters indicated that statement memory in the control group did not differ between "true" and "false" statements, $\Delta G^2(1) = 0.13$, $p = .716$, but was significantly lower for "uncertain" statements, $\Delta G^2\text{s}(1) \geq 3.95$, $ps \leq .047$. In contrast, statement memory was unaffected by feedback type in the interruption group, $\Delta G^2(2) = 2.46$, $p = .292$. Moreover, there was no overall difference in statement memory between the two groups, $\Delta G^2(3) = 6.81$, $p = .078$.

Feedback memory. As expected, feedback memory (*d*-parameters) clearly differed within and between the interruption group and the control group (see Fig. 3). A comparison of memory for the different feedback types within each group replicated the results of Nadarevic and Erdfelder (2013): Both groups showed no significant differences in memory for "true" and "false" feedback, $\Delta G^2\text{s}(1) \leq 2.85$, $ps \geq .092$, whereas memory for "uncertain" feedback was considerably lower than both memory for "true" feedback, $\Delta G^2\text{s}(1) \geq 28.57$, $ps < .001$, and memory for "false" feedback, $\Delta G^2\text{s}(1) \geq 20.11$, $ps < .001$. A comparison of feedback memory between groups showed the expected load effect of the visual discrimination task. That is, feedback memory was generally lower in the interruption group than in the control group, $\Delta G^2(3) = 48.10$, $p < .001$. When investigating this load effect separately for statements with "true," "false," and "uncertain" feedback, we observed the following results: The visual discrimination task at feedback encoding caused lower memory for "true" feedback, $\Delta G^2(1) = 15.22$, $p < .001$, and "false" feedback, $\Delta G^2(1) = 31.27$, $p < .001$, but had no effect on memory for "uncertain" feedback, $\Delta G^2(1) < 0.01$, $p = .986$. Hence, cognitive load impaired memory for "false" *and* "true" feedback.

Guessing. When participants did not know whether a statement was old or new, they showed a strong tendency to guess "new" ($b < .50$) in both experimental groups, $\Delta G^2\text{s}(1) \geq 50.95$, $ps < .001$. Feedback guessing did not differ significantly between recognized and unrecognized statements. That is, *a*-parameters and *g*-parameters could be set equal to each other without a significant decrease in model fit, $\Delta G^2(4) = 3.31$, $p = .507$. Feedback guessing parameters were therefore estimated under the constraint $a = g$ (i.e., $a_{\text{true}} = g_{\text{true}}$, $a_{\text{false}} = g_{\text{false}}$, $a_{\text{uncert.}} = g_{\text{uncert.}}$) to keep the model as parsimonious as possible and thereby minimize standard errors. Cognitive load affected feedback guessing differently, depending on the type of feedback: Participants in the interruption group showed a significantly higher "true" guessing probability as compared with the control group, $\Delta G^2(1) = 9.92$, $p = .002$, but a significantly lower "uncertain" guessing probability, $\Delta G^2(1) = 10.37$, $p = .001$. "False" guessing probabilities did not differ significantly between groups, $\Delta G^2(1) = 3.62$, $p = .057$.

Discussion

Experiment 1 investigated the effect of cognitive load on memory for truth-value feedback for trivia statements. The Spinozan model predicts that cognitive load during feedback processing interferes with the encoding of "false" tags and thus impairs memory for "false" feedback. Memory for "true" feedback, on the other hand, should remain unaffected because the statement's memory representation does not need to be updated as a consequence of the feedback information. In contrast, the Cartesian model predicts impaired memory for "false" *and* "true" feedback, if participants process both types of feedback, because it assumes that cognitive load interferes with both the encoding of "false" tags *and* "true" tags. The results of the present experiment clearly support the Cartesian

[2] Note, however, that this particular pattern of restrictions is not crucial for obtaining the results outlined subsequently. Restricting the detection probabilities for lure statements to zero ($D_{\text{new}} = 0$), as Riefer et al. (1994) did, essentially results in the same pattern of estimates across conditions and does not affect the interpretation of the results. See Appendix B for the parameter estimates based on the $D_{\text{new}} = 0$ restriction.

Table 2 Parameter estimates (with standard errors) of the three-sources MPT model for the control group and the interruption group of Experiment 1

	Statement memory			Feedback memory			Guessing			
	D_{true}	D_{false}	$D_{\text{uncert.}}$	d_{true}	d_{false}	$d_{\text{uncert.}}$	a_{true}	a_{false}	$a_{\text{uncert.}}$	b
Control group	.93 (.01)	.93 (.01)	.90 (.01)	.83 (.02)	.87 (.02)	.43 (.09)	.13 (.02)	.21 (.03)	.66 (.04)	.14 (.04)
Interruption group	.89 (.01)	.91 (.01)	.88 (.01)	.72 (.02)	.70 (.03)	.43 (.05)	.23 (.02)	.28 (.02)	.50 (.03)	.16 (.04)

Feedback guessing parameters were estimated under the constraint $a = g$

model: Memory for both "false" and "true" feedback was lower in the interruption group, in which participants had to perform a visual discrimination task during feedback encoding, than in the control group.

The results of Experiment 1 conceptually replicate Experiment 1 of Nadarevic and Erdfelder (2013) for a dual-task paradigm. Notably, these results are clearly at odds with the findings of Gilbert et al. (1990), who reported evidence compatible with the Spinozan model in a similar dual-task paradigm. However, there are two crucial methodological differences between Gilbert et al.'s Hopi language experiment and our Experiment 1: First, unlike Gilbert and colleagues, we focused on a MPT modeling approach that measures feedback memory corrected for guessing influences. Keep in mind, however, that we found cognitive load effects on memory for "true" and "false" feedback not only when analyzing our data with the MPT model but also when comparing the proportion of correct "true" and "false" attributions by means of CSIMs as Gilbert et al. (Gilbert et al., 1990; Gilbert et al., 1993) did. Thus, our results differ from Gilbert et al.'s findings irrespective of the data analysis approach used. Second, our Experiment 1 and the Hopi language experiment of Gilbert et al. differed with regard to the stimulus materials. Specifically, we used trivia statements, whereas Gilbert and colleagues used nonsense statements, which allegedly represented true or false translations of fictitious Hopi words (e.g., *A monishna is a star*). Hence, possibly, the Spinozan model only applies to artificial materials for which people do not have any stored memory references (Unkelbach & Rom, 2017). To test this possibility, we ran a second experiment that investigated memory for truth-value feedback for statements allegedly representing Hopi–German translations, closely resembling the materials used by Gilbert et al. (1990).

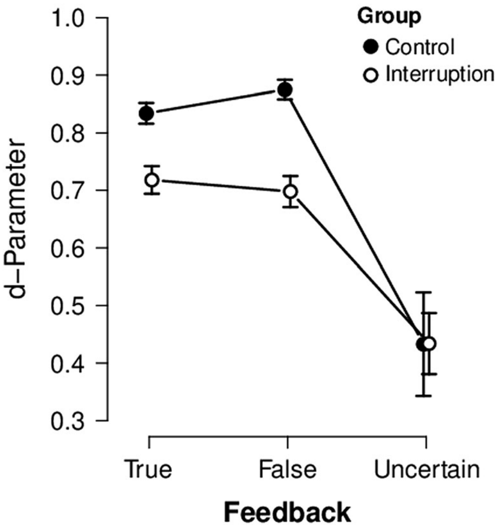

Fig. 3 The *d*-parameter estimates as a function of feedback type and group in Experiment 1. Error bars represent standard errors

Experiment 2

Experiment 2 conceptually replicated Experiment 1. However, instead of studying trivia statements, participants studied Hopi statements along with feedback information on the (alleged) validity of each statement. As in Experiment 1, we assessed feedback memory with two different measures: (a) CSIM and (b) the *d*-parameter of the multinomial model. For CSIM, we expected to replicate the findings reported by Gilbert et al. (1990, Experiment 1) because of the similarity of the experiments, materials, and analyses. That is, we predicted that cognitive load at feedback encoding should reduce the proportion of correct "false" attribution, but not the proportion of correct "true" attributions at test. For the multinomial analyses, we did not set up a specific prediction, but expected to find a pattern of *d*-parameters that is indicative of either the Spinozan model (only memory for "false" feedback decreases under load) or the Cartesian model (memory for both "true" and "false" feedback decreases under load).

Method

As in Experiment 1, a minimum sample size of $N = 60$ was determined a priori. All data exclusions, all manipulations, and all measures employed in the study are reported below.

Participants Ninety participants were recruited at the University of Mannheim and participated for course credit or voluntarily. Four participants were excluded from data analyses for the following reasons: One participant of the interruption group had a very low performance in the visual discrimination task (only 25% correct responses), one participant indicated that she had been familiar with the materials and the hypotheses of the study, and two participants did not fill out the final questionnaire. Thus, the final sample consisted of 86 participants (55 female, 31 male) with a mean age of $M = 23.1$ ($SD = 3.5$) years.

Material We used the 28 statements of the Hopi language experiment of Gilbert et al. (1990) and 37 statements from a similar study of Skurnik (1998) and translated them into German. All statements had the form "an X is a Y" (e.g., a monishna is a star); X was always a fictitious Hopi word, and Y was a German noun. Of the 65 statements, 54 were divided into three stimulus sets, and the remaining 11 statements served as practice or buffer statements, respectively.

Procedure After signing a consent form, the computer instructed participants to imagine that they were traveling through Suriname in South America where they encountered a tribe called Hopi. Similar to Gilbert et al. (1990, Experiment 1), participants were told that they would see statements that supposedly represented Hopi words and their inferred translations followed by feedback on the validity of the translation ("true" vs. "false" vs. "uncertain"). Importantly, the computer instructed participants to memorize statements along with this feedback information. Participants in the interruption group were also informed that during their journey they would oftentimes be phoned by their mother (as indicated by a picture of a mobile phone) or by their friend (as indicated by a picture of a different mobile phone) and were instructed to respond to these phone calls as soon as possible by pressing a left key ("d" for the mother) or right key ("k" for the friend). We implemented eight practice trials in the interruption group in which participants learned to discriminate between the two mobiles phones and their assigned responses. In the practice trials, the pictures of the mobile phones always appeared in the center of the screen, whereas in the later study phase, the pictures appeared randomly on the left or right side of the screen.

The procedure of Experiment 2 was similar to the one of Experiment 1, except for the following changes: Participants studied only 42 statements in the study phase, of which the first three and last three statements served as buffer trials. Each statement was presented for 10 seconds. For each participant, 12 of the 36 target statements were randomly assigned to the feedback "true," 12 to the feedback "false," and 12 to the feedback "uncertain." The final memory test, which directly followed the study phase, comprised 54 statements in total (36 old statements—i.e., 12 from each of the three feedback types—and 18 new ones). Finally, the participants filled out a questionnaire asking whether they had responded seriously and what method they had used to memorize the validity of the statements.

Design As in Experiment 1, feedback type ("true," "false," "uncertain") was manipulated within participants, whereas cognitive load at feedback encoding was manipulated between participants (with $n_1 = 43$ randomly assigned to the interruption group and $n_2 = 43$ to the control group). Moreover, statement sets and distractor set at test were counterbalanced across participants.

Results

Visual discrimination task performance Accuracy in the visual discrimination task was again very high (proportion of correct responses: $M = .98$, $SD = .04$) and unaffected by the feedback type ("true," "false," or "uncertain") that had to be processed while performing the task, $F < 1$.

Memory test performance by condition Replicating Experiment 1, our proxy for statement memory (i.e., the proportion of hits minus the proportion of false alarms) did not significantly differ between the interruption group ($M = .69$, $SD = .16$) and the control group ($M = .66$, $SD = .19$), $t(84) = 0.67$, $p = .502$, $d = 0.15$. We again calculated CSIMs to compare the proportion of correct feedback attributions across groups and feedback types. The descriptive results are displayed in Fig. 4 (for exact CSIMs, see Table 3). A 3 (feedback type: true vs. false vs. uncertain) × 2 (group: interruption vs. control) ANOVA revealed a significant main effect of feedback type on mean CSIMs, $F(2, 168) = 9.71$, $p < .001$, $\eta_p^2 = .10$. Post hoc tests with Bonferroni-adjusted p values showed that this main effect of feedback type was due to higher CSIMs for "true" feedback as compared with "false"

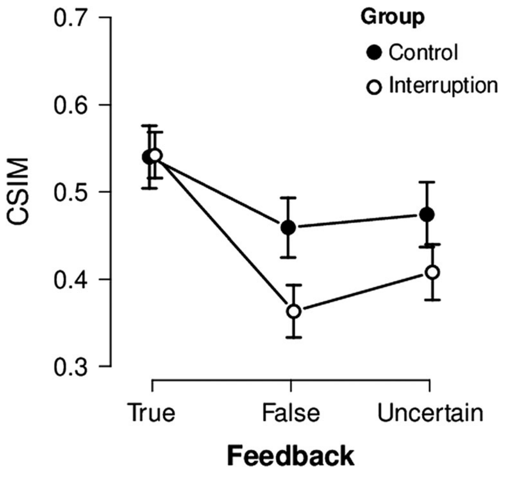

Fig. 4 Mean CSIMs as a function of feedback type and group in Experiment 2. Error bars represent standard errors

Table 3 Mean CSIMs (with standard errors) for the control group and the interruption group of Experiment 2

	Feedback		
	True	False	Uncertain
Control group	.54 (.04)	.46 (.03)	.47 (.04)
Interruption group	.54 (.03)	.36 (.03)	.41 (.03)

feedback, $t(85) = 4.17$, $p < .001$, $d = 0.45$, and as compared with "uncertain" feedback, $t(85) = 3.02$, $p = .010$, $d = 0.33$. CSIMs for "false" and for "uncertain" feedback did not significantly differ, $t(85) = 1.06$, $p = .883$, $d = 0.11$. This time, CSIMs did not differ significantly between groups overall (interruption vs. control group), $F(1, 84) = 3.26$, $p = .075$, $\eta^2_p = .04$, and there was also no Group × Feedback Type interaction, $F(2, 168) = 1.31$, $p = .272$, $\eta^2_p = .02$. However, planned comparisons showed that CSIMs for "false" feedback were significantly lower in the interruption group as compared with the control group, $t(84) = 2.11$, $p = .019$ (one-tailed), $d = 0.46$, whereas CSIMs for "true" feedback did not differ between groups, $t(84) = 0.05$, $p = .520$ (one-tailed), $d = 0.01$. This data pattern is in line with the results of Gilbert et al. (1990), who had also found a selective load effect on "false" attributions, but not on "true" attributions.

Multinomial analyses In order to disentangle statement memory, feedback memory, and guessing processes, data were additionally analyzed with the MPT model specified in Experiment 1. Resembling results for Experiment 1, a likelihood-ratio test indicated that there was no significant model misfit, $G^2(2) = 5.08$, $p = .079$. Parameter estimates of the model are summarized in Table 4.

Statement memory. Similar to Experiment 1, statement memory (i.e., *D*-parameters) in the control group was affected by feedback type, $\Delta G^2(2) = 8.64$, $p = .013$. Pairwise comparisons of the *D*-parameters indicated that participants in the control group remembered statements with "true" feedback significantly better than statements with "uncertain" feedback, $\Delta G^2(1) = 8.35$, $p = .004$. In contrast, they neither showed significant statement memory differences between "true" and "false" statements, $\Delta G^2(1) = 1.07$ $p = .300$, nor between "false" and "uncertain" statements, $\Delta G^2(1) = 3.50$, $p = .062$. In the interruption group, statement memory was unaffected by feedback type, $\Delta G^2(2) = 3.55$, $p = .170$. Again, there was no overall difference in statement memory between the control group and the interruption group, $\Delta G^2(3) = 2.45$, $p = .484$.

Feedback memory. A comparison of feedback memory (*d*-parameters) revealed different data patterns within each group (see Fig. 5). In the control group, the *d*-parameters for the three feedback types differed significantly, $\Delta G^2(2) = 10.04$, $p = .007$. Pairwise comparisons of the *d*-parameters indicated that memory for "true" feedback was significantly better than memory for "false" and "uncertain" feedback, $\Delta G^2\text{s}(1) \geq 6.57$, $ps \leq .010$, whereas the latter did not significantly differ, $\Delta G^2(1) = 0.16$, $p = .688$. In contrast, memory for the three feedback types did not differ significantly in the interruption group, $\Delta G^2(2) = 2.02$, $p = .364$. A comparison of *d*-parameters between groups revealed the following surprising result: We found an cognitive load effect of the visual discrimination task on memory for "true" feedback as indicated by a significantly lower d_{true} parameter in the interruption group as compared with the control group, $\Delta G^2(1) = 5.63$, $p = .018$. However, there was no load effect on memory for "false" or "uncertain" feedback, $\Delta G^2\text{s}(1) \leq 0.75$, $ps \geq .387$. Thus, the feedback memory parameters of the multinomial model revealed a completely different data pattern than the one obtained with the CSIMs, indicating that the CSIM results do not reflect guessing-corrected memory for truth-value feedback.

Guessing. As in Experiment 1, participants of both groups showed a strong tendency to guess "new" ($b < .50$) when they did not know whether a statement was old or new, $\Delta G^2\text{s}(1) \geq 52.21$, $ps < .001$. Moreover, again resembling previous results, feedback guessing parameters *a* and *g* could be set equal to each other, $\Delta G^2(4) = 2.43$, $p = .658$. We therefore estimated feedback guessing parameters under the constraint $a = g$ to compare feedback guessing between groups. Cognitive load affected feedback guessing probabilities differently, depending on the type of feedback: Participants in the interruption group showed a higher "true" guessing probability compared with the control group, $\Delta G^2(1) = 9.66$, $p = .002$, but a significantly lower "false" guessing probability than the control group, $\Delta G^2(1) = 6.02$, $p = .014$. In contrast, the probability

Table 4 Parameter estimates (and standard errors) of the three-sources MPT model for the control group and the interruption group of Experiment 2

	Statement memory			Feedback memory			Guessing			
	D_{true}	D_{false}	$D_{\text{uncert.}}$	d_{true}	d_{false}	$d_{\text{uncert.}}$	a_{true}	a_{false}	$a_{\text{uncert.}}$	b
Control group	.73 (.03)	.69 (.03)	.62 (.02)	.42 (.04)	.23 (.05)	.20 (.06)	.30 (.02)	.33 (.02)	.37 (.02)	.28 (.03)
Interruption group	.73 (.03)	.68 (.03)	.67 (.02)	.25 (.06)	.19 (.04)	.13 (.05)	.38 (02)	.26 (.02)	.35 (.02)	.28 (.03)

Feedback guessing parameters were estimated under the constraint $a = g$

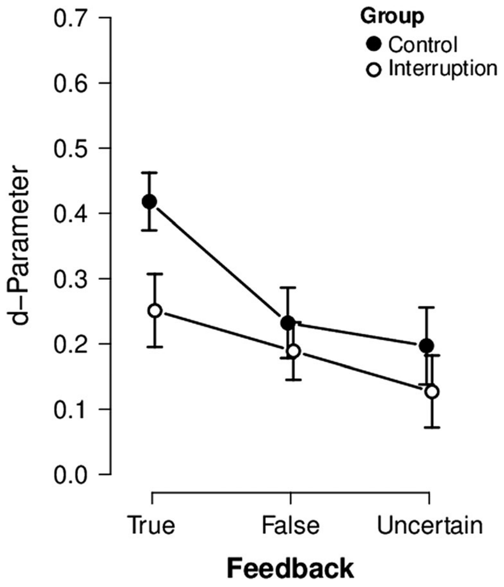

Fig. 5 The *d*-parameter estimates as a function of feedback type and group in Experiment 2. Error bars represent standard errors

to guess "uncertain" did not significantly differ between groups, $\Delta G^2(1) = 0.45$, $p = .501$.

Discussion

Experiment 2 investigated the effect of cognitive load on memory for truth-value feedback for statements of the type investigated by Gilbert et al. (1990, Experiment 1) that allegedly represented translations of a foreign Hopi language. When comparing CSIMs for "true" and "false" feedback between the interruption group and the control group, our results replicated the findings of Gilbert et al.'s Hopi language experiment: The visual discrimination task at feedback encoding reduced the proportion of correct "false" attributions, but not the proportion of correct "true" attributions at test. Although this pattern of results seems to support the Spinozan model for Gilbert-type statements, our MPT model analyses showed otherwise. When using uncontaminated measures of feedback memory (i.e., the *d*-parameters of the MPT model), the visual discrimination task had solely impaired memory for "true" feedback but not memory for "false" feedback, in direct contrast to Gilbert et al.'s (Gilbert et al., 1990; Gilbert et al., 1993) claim. This finding is diametrically opposed to the pattern of the CSIMs and clearly contradicts the Spinozan model. However, because of the lack of a significant load effect on memory for "false" feedback, it is also only partially in line with the predictions of the Cartesian model.

General discussion

The goal of Experiments 1 and 2 was to gain a better understanding of memory for truth-value feedback. According to the Cartesian model, information identified as true is stored along with a "true" tag, and information identified as false is stored with a "false" tag. In contrast, the Spinozan model only distinguishes between untagged and tagged statement representations. The former ones are considered "true," and the latter ones "false." Because the "tagging" process is assumed to require cognitive capacity, Gilbert et al. (1990) tested the two competing models by means of a cognitive load manipulation at truth-value feedback encoding. If the Cartesian model holds, there should be a load effect on memory for "true" and "false" feedback according to their reasoning. In contrast, the Spinozan model predicts a selective load effect on memory for "false" feedback.

Previous studies that aimed to test the Spinozan and the Cartesian model by comparing memory for "true" and "false" feedback either did not disentangle memory and guessing processes properly (Gilbert et al., 1990; Gilbert et al., 1993; Hasson et al., 2005; Richter et al., 2009) or did not investigate the effect of cognitive load on memory for truth-value feedback (Nadarevic & Erdfelder, 2013). The present experiments addressed both of these shortcomings to provide a more thorough test of the two models.

Summary and interpretation of results

The multinomial analyses of the memory test data of Experiment 1 (trivia statements) and Experiment 2 (Hopi language statements) revealed two consistent findings of our cognitive load manipulation: First, cognitive load at feedback encoding increased the probability to guess that a statement had been presented as "true," as indicated by higher "true" guessing parameters in the interruption group as compared with the control group. Second, cognitive load did not selectively impair memory for "false" feedback. In both experiments, the pattern of feedback memory parameters was thus clearly incompatible with the Spinozan model. However, we also did not find unequivocal evidence for the Cartesian model. Because Gilbert et al.'s (Gilbert et al., 1990; Gilbert et al., 1993) specification of the Cartesian model assumes mandatory encoding of "true" and "false" tags in case of sufficient cognitive capacity, the model predicts symmetric cognitive load effects on memory for "true" and "false" feedback. However, only the results of Experiment 1 were in line with this prediction. In contrast, in Experiment 2, cognitive load selectively decreased memory for "true" feedback. Hence, strictly speaking, the results of Experiment 2 are incompatible not only with the Spinozan model but also with the Cartesian model. In other words, neither the Spinozan model nor the Cartesian model can fully account for the findings of Experiment 1 *and* Experiment 2.

What should a model look like that can account for the full pattern of results? First, this model should incorporate both "true" tags and "false" tags, in line with the Cartesian model.

Second, in contrast to the Cartesian model, attaching "true" and "false" tags to encoded information should be optional and context dependent rather than mandatory. More precisely, stored representations of statements might only be tagged as "true" or "false," respectively, when the respective tag is informative in the context defined by the instructions. For instance, in Experiment 2, only Hopi statements with "true" feedback were informative, because only true statements allowed participants to learn the alleged meaning of Hopi words. In contrast, statements with "uncertain" and "false" feedback were equally uninformative in this regard. Hence, it is plausible that participants in Experiment 2 prioritized the encoding of "true" tags, which in turn lead to the selective load effect on memory for "true" feedback in this experiment. In Experiment 1, in contrast, participants were told that their task was to study the trivia statements for a quiz show. In this context, it makes perfect sense to focus on both "true" and "false" feedback because knowledge about false statements would help ruling out false answer options in an anticipated multiple-choice quiz, just as knowledge about true statements would help in recognizing them. This might explain why participants showed symmetrical cognitive load effects on memory for "true" and "false" feedback in Experiment 1.

Future perspectives

In sum, our experiments show that cognitive load during feedback encoding may produce different effects on memory for truth-value feedback, depending on the experimental context (i.e., the presented statements and the cover story). These results favor a context-dependent tagging model over the Spinozan model and the Cartesian model. However, because the proposed context-sensitive tagging model incorporates both "true" and "false" tags, we believe that it is much closer to the spirit of the Cartesian model than to that of the Spinozan model. The only differece to the Cartesian model is that it does not rest on a mandatory tagging assumption. Apparently, in some contexts (e.g., in Experiment 2) our cognitive system places more emphasis on the encoding of "true" tags than on "false" tags (or vice versa), whereas in other contexts (e.g., in Experiment 1) it does not prioritize one of the two tags.

An essential next step is thus to gain a better understanding of the interplay of factors that determine the encoding of "true" tags and "false" tags, respectively. For this purpose, different characteristics of the presented statements and the study context should be investigated in more detail, and additional variables should be taken into account. For instance, work by Street and colleagues (Street & Richardson, 2014; Street, Bischof, Vadillo, & Kingstone, 2016) suggests that true/false evaluations can be influenced by people's beliefs about the base rate of true and false information. However, it is unclear so far whether such beliefs (e.g., "word definitions are generally true" or "there are many dubious social media postings") also moderate the encoding of "true" tags and "false" tags, respectively. Likewise, it would be interesting to investigate memory for truth-value feedback in contexts in which lying is the norm, that is, in contexts with a high base rate of "false" statements. Finally, future studies should also examine possible differences between intentional versus incidental feedback learning on the encoding of "true" and "false" tags.

For the sake of ecological validity, we argue that the proposed research questions should be investigated with meaningful statements instead of nonsense statements. This undertaking is particularly important and practically relevant in times where statement evaluations are susceptible to fake news and alternative facts.

Acknowledgments This work was supported by a University of Mannheim autonomy grant. We thank Johanna Heckeroth, Denise Meyer, Jana Ritschel, and Vera Vogel for collecting data of Experiment 1. Moreover, we thank Vanessa Gottschall for her dedicated help with setting up and collecting data for Experiment 2.

Open practices statement All materials and data sets for the two experiments are available via the Open Science Framework (OSF): https://osf.io/ac4rt/

Appendix 1

Multinomial processing tree (MPT) models (for a review, see Erdfelder et al., 2009) are stochastic models that are based on assumptions about the interplay of various latent cognitive processes that presumably underlie an observed behavioral response in a particular experimental paradigm (e.g., an "old" response in a recognition test). The assumed interaction of the proposed underlying processes (e.g., memory processes and guessing processes) can be illustrated as a processing tree. A processing tree links a particular test stimulus with all possible response options to this stimulus by specifying different sequences of cognitive processes that may lead to the behavioral responses. The probabilities of relevant cognitive processes taking place or not are formalized as model parameters that are bound to individual branches of the processing tree and reflect transition probabilities between cognitive states.

The feedback memory model described in the main text—an adapted version of the three sources model of Riefer et al. (1994)—distinguishes four different stimulus types in the memory test: statements with "true" feedback, statements with "false" feedback, statements with "uncertain" feedback, and "new" statements. For each of the four stimulus types, a separate processing tree is proposed, each making specific assumptions about the possible cognitive processes that may lead to the four response options "true," "false," "uncertain," and "new." Four stimulus types times four response options result in overall 16 different possible outcome events that form the empirical basis of the MPT model. These events are determined by memory and guessing processes represented by the following model parameters: *D* (probability of statement recognition or lure detection, respectively), *d* (probability of feedback recognition), *b* (probability of guessing "old" in case of recognition uncertainty), a_i (probability of guessing feedback *i* for recognized statements), and g_i (probability of guessing feedback *i* for unrecognized statements).

In the following, we will explain the MPT model illustrated in Fig. 1 in the main text using the example of a statement with "true" feedback. When presented with such a statement in the memory test, a participant either recognizes this statement as old with probability D_{true} or does not recognize the statement with the complementary probability $1 - D_{\text{true}}$. If the statement is recognized, the feedback "true" may also be recognized with probability d_{true} or it is not recognized with the complementary probability $1 - d_{\text{true}}$. If both the statement and the feedback is recognized, the participant responds "true" at test accordingly. If the statement is recognized but the feedback is not, the participant must guess the truth-value feedback presented along with the statement. Specifically, the participant guesses with probability a_{true} that the statement was presented as "true," with probability a_{false} that the statement was presented as "false," and with probability $a_{\text{uncert.}}$ that the statement was presented as "uncertain." If the statement is not recognized, the participant is in a state of recognition uncertainty. In this case, the participant either guesses with probability *b* that the statement is old or with probability 1 − *b* that the statement is new. In the latter case, the participant responds "new." However, in case of an "old" guess, the feedback information has to be guessed as well. Hence, the participant guesses "true" with probability g_{true}, "false" with probability g_{false}, and "uncertain" with probability g_{uncert}.

The same logic also applies to statements with "false" feedback and those with "uncertain" feedback. For the "new" statements (i.e., the lures in the memory test), the D_{new} parameter does not reflect statement recognition, but lure detection—that is, the probability to detect that a new statement was not presented in the study phase. In contrast, $1 - D_{\text{new}}$ reflects the probability that a lure remains undetected. In this case, participants are in the state of uncertainty, which again leads to the same guessing processes as recognition uncertainty for old target statements. By implication, because new statements were not presented in the study phase—and thus did not receive truth-value feedback—the processing tree for the new statements does not include a feedback memory parameter (*d*).

The MPT model illustrated in Fig. 1 can be translated into model equations that specify the probabilities of the 16 outcome events in the memory test as a function of the model's parameters. For instance, *p*("true"/false) denotes the probability of responding "true" to a statement presented with *false* feedback. This leads to the following 16 model equations (note that parameter indices are abbreviated below; t = true, f = false, u = uncert., n = new):

(1) $p(\text{"true"/true}) = D_t \times d_t + D_t \times (1 - d_t) \times a_t + (1 - D_t) \times b \times g_t$
(2) $p(\text{"false"/true}) = D_t \times (1 - d_t) \times a_f + (1 - D_t) \times b \times g_f$
(3) $p(\text{"uncert."/true}) = D_t \times (1 - d_t) \times a_{u.} + (1 - D_t) \times b \times g_u$
(4) $p(\text{"new"/true}) = (1 - D_t) \times (1 - b)$
(5) $p(\text{"true"/false}) = D_f \times (1 - d_f) \times a_t + (1 - D_f) \times b \times g_t$
(6) $p(\text{"false"/false}) = D_f \times d_f + D_f \times (1-d_f) \times a_f + (1-D_f) \times b \times g_f$
(7) $p(\text{"uncert."/false}) = D_f \times (1 - d_f) \times a_u + (1 - D_f) \times b \times g_u$
(8) $p(\text{"new"/false}) = (1 - D_f) \times (1 - b)$
(9) $p(\text{"true"/uncert.}) = D_u \times (1 - d_u) \times a_t + (1 - D_u) \times b \times g_t$
(10) $p(\text{"false"/uncert.}) = D_u \times (1 - d_u) \times a_f + (1 - D_u) \times b \times g_f$
(11) $p(\text{"uncert."/uncert.}) = D_u \times d_u + D_u \times (1 - d_u) \times a_u + (1 - D_u) \times b \times g_u$
(12) $p(\text{"new"/uncert.}) = (1 - D_u) \times (1 - b)$
(13) $p(\text{"true"/new}) = (1 - D_n) \times b \times g_t$
(14) $p(\text{"false"/new}) = (1 - D_n) \times b \times g_f$
(15) $p(\text{"uncert."/new}) = (1 - D_n) \times b \times g_u$
(16) $p(\text{"new"/new}) = D_n + (1 - D_n) \times (1 - b)$

With these model equations, it is possible to estimate the model's parameters for a given data set of response frequencies by means of an iterative maximum likelihood estimation algorithm (EM algorithm; Hu & Batchelder, 1994), provided that the model is identifiable. Bayen et al. (1996) noted that two-high-threshold MPT models (such as the MPT model described above) are not identifiable without restricting the value of the D_{new}-parameter. For this reason, they suggested to equate D_{new} with at least one of the other *D*-parameters. This parameter restriction is in line with the empirically well-established mirror effect—the symmetrical increase (or decrease) of hits and correct rejections (Glanzer et al., 1993). Indeed, a validation study by Bayen et al. (1996) provided convincing evidence for the superiority of the two-high-threshold model with restricted D_{new}-parameter compared with one-high-threshold and low-threshold models. In our two experiments, we restricted the value of the D_{new}-parameter to the value of the $D_{\text{uncert.}}$-parameter within each group (see Bell et al., 2010, for an equivalent restriction). Importantly, this restriction did not result in model misfit.

Model fit of MPT models can be assessed by means of the goodness-of-fit statistic G^2, which is asymptotically χ^2-distributed if the model holds (Read & Cressie, 1988). The G^2-test compares the model's predicted response frequencies with the observed response frequencies. The degrees of freedom for this test correspond to the number of independent outcome events minus the number of freely estimated parameters. Significant discrepancies between the predicted and observed frequencies indicate model misfit. However, model fit is not the only criterion for the evaluation of a model's validity. A valid MPT model is characterized by the fact that each of its parameters responds to targeted experimental manipulations in a predictable manner. This criterion is met by source-monitoring MPT models—such as the model described above and used in our experiments—as has been successfully demonstrated in several validation studies (Bayen et al., 1996; Riefer et al., 1994). Most importantly for our purposes, the *d*-parameter of such models is a much more accurate measure of source memory (or feedback memory, respectively) than proxy measures such as SIM and CSIM, which may be confounded by item memory and guessing processes (for a review, see Bröder & Meiser, 2007).

Appendix 2

Table 5 Mean parameter estimates (with standard errors) of the one-high-threshold three sources MPT model for Experiments 1 and 2

	Experiment 1			Experiment 2		
	Control group	Interruption group	$\Delta G^2(1)$	Control group	Interruption group	$\Delta G^2(1)$
D_{true}	.94 (.01)	.91 (.01)	5.84*	.78 (.02)	.79 (.02)	0.05
D_{false}	.94 (.01)	.93 (.01)	0.89	.75 (.02)	.75 (.02)	0.01
$D_{\text{uncertain}}$	.91 (.01)	.90 (.01)	0.24	.69 (.02)	.74 (.02)	1.83
d_{true}	.83 (.02)	.71 (.02)	16.97***	.39 (.04)	.24 (.05)	5.77*
d_{false}	.87 (.02)	.69 (.03)	33.53***	.21 (.05)	.17 (.04)	0.41
$d_{\text{uncertain}}$	.42 (.09)	.42 (.05)	0.00	.18 (.05)	.11 (.05)	0.71
a_{true}	.13 (.02)	.23 (.02)	9.92**	.30 (.02)	.38 (.02)	9.66**
a_{false}	.21 (.03)	.28 (.02)	3.62	.33 (.02)	.26 (.02)	6.02*
$a_{\text{uncertain}}$	.66 (.04)	.50 (.03)	10.37**	.37 (.02)	.35 (.02)	0.45
b	.02 (.00)	.02 (.00)	0.38	.11 (.01)	.09 (.01)	1.04

Model fit in Experiment 1: G^2 (2) = 5.39, p = .067. Model fit in Experiment 2: G^2 (2) = 5.06, p = .080. Feedback guessing parameters were estimated under the constraint $a = g$

*$p < .05$. **$p < .01$. ***$p < .001$

References

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. https://doi.org/10.1037/0278-7393.22.1.197

Bell, R., Buchner, A., & Musch, J. (2010). Enhanced old–new recognition and source memory for faces of cooperators and defectors in a social-dilemma game. *Cognition, 117,* 261–275. https://doi.org/10.1016/j.cognition.2010.08.020

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Erlbaum.

Descartes, R. (1985). Principles of philosophy. In J. Cottingham, R. Stoothoff, & D. Murdoch (Eds.), *The philosophical writings of Descartes* (Vol. 1, pp. 177–291). Cambridge, UK: Cambridge University Press. (Original work published 1644)

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. https://doi.org/10.1027/0044-3409.217.3.108

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

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. https://doi.org/10.1037/0022-3514.59.4.601

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. https://doi.org/10.1037/0022-3514.65.2.221

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567. https://doi.org/10.1037/0033-295X.100.3.546

Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C. (1999). Not propositions. *Journal of Cognitive Systems*

Research, 1, 19–33. https://doi.org/10.1016/S1389-0417(99)00004-2

Grice, H. P. (1989). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 22–40). Cambridge, MA: 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. https://doi.org/10.1111/j.0956-7976.2005.01576.x

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59,* 21–47. https://doi.org/10.1007/BF02294263

Hunt, E. (2017). 'Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news. *The Guardian.* Retrieved from https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news.

Isberner, M.-B., & Richter, T. (2013). Does validation during language comprehension depend on an evaluative mindset? *Discourse Processes, 51,* 7–25. https://doi.org/10.1080/0163853X.2013.855867

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the "Who said what?" paradigm. *Journal of Personality and Social Psychology, 75,* 1155–1178. https://doi.org/10.1037/0022-3514.75.5.1155

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

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

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

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

Piest, B. A., Isberner, M.-B., & Richter, T. (2018). Don't believe everything you hear: Routine validation of audiovisual information in children and adults. *Memory & Cognition, 46,* 849–863. https://doi.org/10.3758/s13421-018-0807-7

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York, NY: Springer-Verlag.

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,* 538–558. https://doi.org/10.1037/a0014038

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. https://doi.org/10.1037/0278-7393.20.3.680

Skurnik, I. W. (1998). *Metacognition and the illusion of truth* (Unpublished doctoral dissertation). Princeton University, Princeton, NJ.

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

Street, C. N. H., Bischof, W. F., Vadillo, M. A., & Kingstone, A. (2016). Inferring others' hidden thoughts: Smart guesses in a low diagnostic world. *Journal of Behavioral Decision Making, 29,* 539–549. https://doi.org/10.1002/bdm.1904

Street, C. N. H., & Kingstone, A. (2016). Aligning Spinoza with Descartes: An informed Cartesian account of the truth bias. *British Journal of Psychology, 108,* 453–466. https://doi.org/10.1111/bjop.12210

Street, C. N. H., & Richardson, D. C. (2014). Lies, damn lies, and expectations: How base rates inform lie–truth judgments. *Applied Cognitive Psychology, 29,* 149–155. https://doi.org/10.1002/acp.3085

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

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. https://doi.org/10.1037/0278-7393.33.2.443

Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (2013). Validating the truth of propositions: Behavioral and ERP indicators of truth evaluation processes. *Social Cognitive and Affective Neuroscience, 8,* 647–653. https://doi.org/10.1093/scan/nss042