Does validation during language comprehension depend on an evaluative mindset?

Maj-Britt Isberner and Tobias Richter

University of Kassel

*Word Count:* 6,059

accepted for publication in the journal *Discourse Processes* (2013)

Acknowledgments

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, grant RI 1100/4-2). We would like to thank Benjamin Piest for his help in data collection.

Corresponding author:

Maj-Britt Isberner

University of Kassel, Department of Psychology

Holländische Str. 36-38

34109 Kassel, Germany

E-mail: maj-britt.isberner@uni-kassel.de

Telephone: +49-561 804 1905

Fax: +49-561 804 3586
**Abstract**

Whether information is routinely and nonstrategically evaluated for truth during comprehension is still a point of contention. Previous studies supporting the assumption of nonstrategic validation have used a Stroop-like paradigm in which participants provided yes/no judgments in tasks unrelated to the truth or plausibility of the experimental sentences. Other studies using a nonevaluative task failed to support this assumption. This leaves open the possibility that validation is conditional on an evaluative mindset of the reader. In the present study, we investigated this question directly by using a nonevaluative probe task. Participants responded to the probe words "true" or "false" with two different keys after reading true or false sentences for comprehension. Results provide evidence for routine validation even when it is not encouraged by the task but they also suggest that semantic processing is critical for validation to occur. These results can be taken as evidence for a close connection between validation and comprehension rather than validation being a goal-dependent process.
Whether information is routinely evaluated for validity (or truth) during language comprehension is still a point of contention. A widely accepted view is that validation – that is, computing truth values or plausibility based on relevant world knowledge – is a strategic, optional process which is subsequent to comprehension (Gilbert, 1991; Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993; Herbert & Kübler, 2011). Based on this idea, two-step models of comprehension and validation either assume that comprehension proceeds without any evaluative component (e.g., Connell & Keane, 2006) or that the linguistic input is by default initially accepted as true and can only effortfully be “unbelieved” at a later point (e.g., Gilbert et al., 1990, 1993). However, many psycholinguistic studies implicitly or explicitly call into question the conceptualization of comprehension and validation as nonoverlapping stages of information processing. In fact, validation is often utilized in psycholinguistic studies to measure comprehension in the first place: Sentence verification has been a popular tool to assess the time it takes to understand a sentence, thus allowing conclusions regarding the organization of semantic memory (e.g., Kintsch, 1980; Kounios, Osman, & Meyer, 1987). Similarly, readers’ ability to detect inconsistencies with their knowledge or with prior text information is frequently utilized to study the kinds of knowledge or portions of prior discourse that are accessed during comprehension, and the time courses of their activation (Fischler, Childers, Achariyapaopan, & Perry, 1985; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Nieuwland & van Berkum, 2006; O’Brien, Rizzella, Albrecht, & Halleran, 1998; Rayner, Warren, Juhasz, & Liversedge, 2004; van Berkum, Zwitserlood, Hagoort, & Brown, 2003), as well as how (quickly) particular syntactic structures are interpreted (e.g., Nieuwland & Kuperberg, 2008; Pickering & Traxler, 1998; Speer & Clifton, 1998; Staub, Rayner, Pollatsek, Hyönä, & Majewski, 2007; Traxler & Pickering, 1996; Van Gompel, Pickering, & Traxler, 2001).
A prominent example is the study of memory-based processes: These have been demonstrated by introducing an inconsistency with prior information into a text (e.g., *Mary is a vegetarian. [...] She orders a cheeseburger*), which generally results in longer reading times on the inconsistent sentence (e.g., Albrecht & O’Brien, 1993; O’Brien et al., 1998). This finding is taken as evidence for the reactivation of prior text information by memory-based processes, but it also shows how routinely readers detect inconsistencies during comprehension. In a similar vein, studies on how and when linguistic input is related to the wider discourse have used inconsistencies, which generally elicit an elevated N400 event-related potential (ERP) component, to show the immediate integration of new information into its context (Nieuwland & van Berkum, 2006; van Berkum, Zwitserlood, Hagoort, & Brown, 2003). Following a similar logic, Hagoort et al. (2004) provided evidence for simultaneous integration of semantic knowledge and world knowledge in language comprehension by showing that semantic violations (*Dutch trains are sour*) and world knowledge violations (*Dutch trains are white*) elicit an N400 of similar time course and magnitude compared to correct control information (*Dutch trains are yellow*). This not only calls the traditional distinction between semantic knowledge (relevant for comprehension) and world knowledge (relevant for validation) into question (see also Jackendoff, 2002), but also clearly speaks “against a nonoverlapping two-step interpretation procedure in which first the meaning of a sentence is determined, and only then is its meaning verified in relation to our knowledge of the world” (Hagoort et al., 2004, p. 440).

Consistent with this notion, Singer (2006) proposed that "Memory-based processes afford the verification of the current text constituent." (p. 587). Essentially, this implies that the information which is passively activated for the comprehension of incoming information concurrently allows its validation. Singer and colleagues have shown that this is true for so-called bridging inferences, which causally link sentences (e.g., Singer, 1993; Singer, Halldorson, Lear, & Andrusiak, 1992): When participants read sentences such as *Dorothy*
poured the bucket of water on the fire. The fire went out or Dorothy poured the bucket of water on the fire. The fire grew hotter, which imply a causal relationship that is either consistent or inconsistent with general world knowledge, they are subsequently faster to answer the question Does water extinguish fire? than after reading sentences which did not imply a causal relationship (i.e., Dorothy placed the bucket of water by the fire. The fire went out / grew hotter). This suggests that causal bridging inferences are not only routinely generated during comprehension but at the same time validated against relevant world knowledge.

All of these results suggest a close connection or even an overlap between comprehension and validation, which is rarely explicitly addressed (but see Fischler & Bloom, 1980; Fischler et al, 1983; Isberner & Richter, 2013; Murray & Rowan, 1998; Richter, Schroeder, & Wöhrmann, 2009; Singer, 2006; West & Stanovich, 1982, for exceptions). Moreover, they suggest that validation is a routine component of comprehension under normal reading conditions – that is, without the explicit instruction to validate incoming information in relation to world knowledge or prior discourse. In line with this idea, Richter et al. (2009) and Isberner and Richter (2013) found evidence for nonstrategic validation in terms of Stroop-like stimulus response compatibility effects (Stroop, 1935). In these experiments, participants read sentences presented word-by-word and were prompted to respond with a positive or negative response (independent of validity) at varying points during sentence presentation. In experimental items, which varied in validity (e.g., Perfume contains scents/Soft soap is edible) or plausibility (e.g., Frank has a broken pipe/leg. He calls the plumber), the prompt always appeared immediately after the end of the sentence. With this paradigm, positive and negative responses in an orthographical judgment task (Is the word spelled correctly? Isberner & Richter, 2013; Richter et al., 2009) and a color judgment task (Did the word change color? Isberner & Richter, 2013) were shown to be slower when they were incongruent with the truth value or plausibility of the sentence read prior to responding.
(i.e., a positive/‘correct’/‘yes’ response after a false/implausible sentence or a negative/‘incorrect’/‘no’ response after a true/plausible sentence) than when they were congruent. This suggests that readers cannot ignore validity or plausibility even when it is irrelevant to the experimental task (irrelevant stimulus-response compatibility). Based on these findings, Richter et al. (2009) suggested that comprehension comprises an epistemic monitoring process, which detects inconsistencies with easily accessible prior knowledge and thus protects the mental representation (situation model; Zwaan & Radvansky, 1998) from contamination with false or implausible information (Schroeder, Richter, & Hoever, 2008).

However, the generality of these findings has recently been called into question by a study by Wiswede, Koranyi, Müller, Langner, and Rothermund (2012) which seems to show that the Stroop-like compatibility effect is conditional on an evaluative mindset. Although validity was irrelevant in the orthographical task used by Richter et al., Wiswede et al. noted that the correct/wrong orthography decision may have induced an evaluative mindset which may have encouraged evaluation of the stimuli for validity. Thus, Wiswede et al. attempted to show that effects of automatic validation hinge on an evaluative mindset of the reader.

For this purpose, they asked participants to read obviously true and false sentences (e.g., *Africa is a continent* or *Saturn is not a planet*) presented word-by-word using Rapid Serial Visual Presentation. After the presentation of each sentence, which was followed by a blank screen of 1500 ms, a prompt was presented signaling which of two randomly intermixed tasks participants had to perform on the current trial. One of the two tasks was a simple probe word identification task in which participants had to respond to a probe which either read “TRUE” or “FALSE” with the associated response key. Importantly, the probe was independent of the actual truth value of the sentence; that is, it matched the truth value on a random half of the trials (e.g., *Africa is a continent* – TRUE) and on the other half it did not (e.g., *Africa is a continent* – FALSE). Participants only had to respond to the probe, regardless of whether it matched the truth value of the previous sentence or not.
To induce an evaluative or nonevaluative mindset, Wiswede et al. (2012) intermixed this probe task randomly with a second task, which differed between two groups of participants. In the evaluative mindset group, the second task was a truth evaluation task, in which the participants were prompted to decide about the truth value of the sentence (e.g., *Africa is a continent* – True or false?). Thus, evaluating the sentences regarding their truth value was encouraged in this group because it was useful on one of the two tasks (i.e., on a random half of the trials). In the nonevaluative mindset group, the second task was a sentence comparison task. Participants were shown a sentence which was either the same as the one they had read before (e.g., *Africa is a continent* – *Africa is a continent*) or a slightly different one (e.g. *Africa is a continent* – *Africa is a planet*) and had to indicate whether the second sentence was the same as the first (“Is this the sentence that you’ve just seen?”). Thus, in this task, evaluating the truth value of the sentences was not beneficial for completing either of the two tasks. Wiswede et al. assumed that as participants did not know which of the two tasks they would have to perform on each trial until the response prompt appeared, the demands of the second task (or the mindset induced by the second task) would affect sentence processing in the probe task as well, which should result in performance differences in this task between the two groups. Specifically, they expected that only the evaluative mindset group would exhibit interference if the probe (TRUE or FALSE) did not match the actual truth value of a sentence, whereas the nonevaluative mindset group would not spontaneously evaluate the sentences (given that it was not required by either of the two tasks), and thus not show any interference.

In line with this prediction, Wiswede et al. (2012) found a compatibility effect in terms of a Truth X Probe interaction as well as ERP evidence for validation in the evaluative mindset group, but not in the nonevaluative mindset group. Thus, they concluded that validation is conditional on an evaluative mindset of the reader.
In the present study, we call this conclusion into question. We argue that the two different tasks Wiswede et al. (2012) intermixed with the probe task differed not only in whether they encouraged evaluation, but more generally regarding the depth of semantic processing that was required. Consider, for example, that the sentence comparison task could also be performed in a foreign language at a purely perceptual level (even though it must be noted that the first sentence was presented with Rapid Serial Visual Presentation while the second sentence was presented all at once). This idea is supported by the fact that effects of semantic mismatches on the amplitude of the N400, which is associated with semantic processing (e.g., Kutas & Federmeier, 2011), were significantly reduced in the nonevaluative mindset group ($\eta_p^2 = .84$ in the evaluative mindset group vs. $\eta_p^2 = .36$ in the nonevaluative mindset group). Naturally, if the depth of semantic processing is reduced, the effectiveness of validation will be impaired as well. Therefore, in order to investigate the conditionality of validation, it would be more appropriate to use a task which, while still requiring an adequate depth of processing of the stimuli, does not explicitly encourage validation. Our main goal in the present study was to find such a task, and to test whether it produces compatibility effects as a function of validity and required responses, which are consistent with routine validation. This finding would support the idea that merely understanding a sentence by default entails its validation, provided that a reader has easily accessible knowledge which allows assessing its validity.

To test this assumption, we used the probe task by Wiswede et al. (2012) but combined it with comprehension questions which did not require validation of the sentences – namely, whether or not a particular sentence involved an animate object. In this way, our task ensured comprehension of the sentences without encouraging validation.

Another open question is whether the reported compatibility effects reflect facilitation for compatible conditions, interference for incompatible conditions, or both. Thus, a second goal of our study was to address this question. As has been noted in the Stroop literature, in
order to study interference and facilitation effects, it is necessary to use an adequate neutral condition (e.g., MacLeod, 1991). For the purpose of the present study, our idea was to use as a neutral condition stimuli for which participants have no (or little) knowledge which would allow them to assess the validity of the sentences (e.g., *Toothpaste contains sulfur*). Thus, these sentences should not create interference with or facilitation of positive and negative responses because participants do not possess easily accessible knowledge required for assessing validity.

If our assumptions hold, we should thus find a compatibility effect in the response latencies but only for items for which participants have high knowledge. Thus, we expect a three-way interaction of knowledge, validity, and required response, driven by a two-way interaction of the latter two variables (compatibility effect) emerging only in the high knowledge condition. Moreover, compatibility effects may show not only in the response latencies, but also in the error rates (e.g., Richter et al., 2009).

**Method**

**Participants**

Participants were 42 students of various subjects at the University of Kassel. The data of 9 nonnative speakers of German were excluded from the analysis. The average age of the 33 remaining participants (21 female) was 23.3 years ($SD = 2.9$; range 19-33 years). Participants provided informed consent at the beginning of the experiment and were reimbursed with 6 € after its completion.

**Stimulus material**

The stimuli were valid (true) and invalid (false) sentences of the structure “[a] [concept noun] [is/has/causes/contains] [a] [concept noun/adjective].” - for example *Perfume contains scents* (the actual stimuli were in German; e.g., *Parfüm enthält Duftstoffe*). The materials were taken from the study by Richter et al. (2009, Experiment 4), having already
been normed for validity and knowledge. The 12 participants in that norming study were asked to indicate for 288 items in total (144 true and 144 false) whether the sentences were true or false and how certain they were in their judgment on a 6-point scale ranging from 1 (very uncertain) to 6 (very certain). This norming study allowed for grouping the items according to knowledge (high knowledge: high agreement between participants and high average judgment certainty vs. low knowledge: low agreement between participants and low average judgment certainty).

**Experimental items.** From this pool of items, 96 experimental items were drawn. Half of these were associated with high knowledge, i.e., the true items had consistently been judged as true (mean agreement: 100%) with high judgment certainty ($M = 5.73, SD = 0.18$), e.g., *Perfume contains scents*, and the false items had consistently been judged as false (mean agreement: 98%) with high judgment certainty ($M = 5.78, SD = 0.10$), e.g., *Soft soap is edible*. These items were identical to the 48 experimental items used in Experiment 4 by Richter et al.

In addition, we used 24 true and 24 false items for which participants in the norming study had exhibited low knowledge (e.g., *Krypton is a noble gas* or *Toothpaste contains sulfur*); that is, the truth value of these items had been judged inconsistently in the norming study (mean agreement: 56%) and with on average low judgment certainty ($M = 2.83, SD = 0.95$). These were part of the filler items in Experiment 4 by Richter et al. (2009) but were used as experimental items in the present study.

**Filler items.** For the filler trials, we used 56 additional items from the Richter et al. (2009) material, of which 32 were associated with low knowledge and 24 were associated with high knowledge. Of each of these, half were true and half were false.

**Procedure**

Participants were tested in a computer lab in groups of up to 5 people. They were asked to rest the index fingers of their left and right hand on the two response keys throughout the experiment, and to respond as fast and as accurately as possible. All stimuli were
presented in an individually randomized order. Every 40 trials, participants were allowed to take a short break. The first 8 trials were practice trials, after which participants had the opportunity to ask questions before starting the actual experiment.

On all trials, the stimuli were presented word-by-word on a computer screen using Rapid Serial Visual Presentation with a fixed rate of 300 ms per word; all words were presented in black font (Arial, approximate height 1 cm) against a white background. Every trial was followed by a blank screen presented for 1000 ms. Figure 1 displays the trial structure of experimental and filler trials.

**Experimental trials.** Our Stroop-like task combined the procedure employed by Richter et al. (2009) with the probe identification task used by Wiswede et al. (2012). In the experimental 96 trials, the probe "**Richtig**" ("True") or "**Falsch**" ("False") appeared after the third and final word of the stimulus sentence, prompting the participants to respond with the corresponding key ("k" for the "True" probe, "d" for the "False" probe). Half of the trials were presented with the “True” probe, the other half with the “False” probe. The probe was presented orthogonally to the validity of the sentence; that is, it was independent of validity and thus either matched or mismatched the actual truth value of the sentence. Importantly, participants were only required to identify the probe word and press the corresponding key, regardless of whether the probe matched or mismatched the validity of the sentence. The probe remained on the screen until the participant provided a response.

**Filler trials.** In the 56 filler trials, as in the study by Richter et al. (2009), the probe appeared after the first word (28 trials) or the second word (28 trials) of the sentence, in order to make the appearance of the probe less predictable and the goal of the study (investigating effects of the match or mismatch of the probe with the actual truth value of the sentence) less transparent. Half of the filler trials were presented with the “True” probe, whereas the other half were presented with the “False” probe.
Comprehension questions. In each of the 28 filler trials on which the prompt appeared after the first word, a comprehension question was presented immediately after the sentence which – crucially – required comprehension but not validation of the sentence (see Figure 1). Thus, comprehension was ensured without inducing an evaluative mindset. Specifically, participants were asked to indicate whether the sentence had referred to an animate object; five of the comprehension questions required a yes-response. Participants were informed before the experiment that they would be asked comprehension questions and were instructed to process the content of the sentences to be able to answer these questions. To make sure that participants understood the importance of reading for comprehension before starting the actual experiment, 2 of the 8 practice trials comprised comprehension questions.

Design

The design was a 2 (knowledge: high vs. low) X 2 (validity: valid vs. invalid) X 2 (required response: positive vs. negative) within-subjects design. The assignment of the probes “True” and “False” to the stimuli was counterbalanced via two item lists. Response latencies and accuracy in the probe task were recorded as dependent variables.

Results

Type-I error probability was set at .05 for all hypothesis tests. ANOVAs were conducted for repeated measurements with both participants ($F_1$, by-subjects) and items ($F_2$, by-items) as the source of random variance. The reported means and standard errors were computed with subjects as the units of observation.

Comprehension questions

The imbalance between items requiring yes- and no- responses was due to the fact that we used the already normed material from Richter et al. (2009), which had not been constructed to be balanced regarding animacy of the objects. Given that participants received the questions in a random order, however, we think that it is unlikely that this induced a response strategy in the reader.
On average, participants answered 85.9 % ($SD = 7.8$ %) of the comprehension questions correctly.

**Response latencies**

Response latencies were included for correct responses (94.9% of the responses in experimental trials). Latencies deviating more than three standard deviations from either the subject or item mean (2.1% of all correct latencies) were treated as outliers and removed from the data set.

The full data for the response latencies are displayed in Figure 2. In the overall 2 X 2 ANOVA, we found main effects of all three independent variables: Items for which participants had high knowledge ($M = 700, SE = 43$) were responded to faster than items for which participants had low knowledge ($M = 720, SE = 46$), $F(1, 32) = 5.92, p < .05, \eta_p^2 = .16, F_2(1, 92) = 7.05, p < .01, \eta_p^2 = .07$; valid items ($M = 693, SE = 41$) were responded to faster than invalid items ($M = 727, SE = 48$), $F_1(1, 32) = 12.94, p < .01, \eta_p^2 = .29, F_2(1, 92) = 8.47, p < .01, \eta_p^2 = .08$; and positive responses ($M = 695, SE = 43$) were given faster than negative responses ($M = 725, SE = 46$), $F_1(1, 32) = 5.44, p < .05, \eta_p^2 = .15, F_2(1, 92) = 5.29, p < .05, \eta_p^2 = .05$. In addition, there was a two-way interaction between validity and required response in the by-subjects analysis, $F_1(1, 32) = 4.69, p < .05, \eta_p^2 = .13, F_2(1, 92) = 2.12, p = .15$. However, this interaction was qualified by a three-way interaction between all three independent variables, $F_1(1, 32) = 4.26, p < .05, \eta_p^2 = .12, F_2(1, 92) = 6.07, p < .05, \eta_p^2 = .06$.

We followed up on this finding by running separate analyses for the high and low knowledge conditions.

**High versus low knowledge.** Whereas in the low knowledge condition, there were no significant effects, all $p > .10$, in the high knowledge condition there were significant main effects of validity, $F_1(1, 32) = 12.31, p < .01, \eta_p^2 = .28, F_2(1, 46) = 16.27, p < .001, \eta_p^2 = .26$, with valid items ($M = 675, SE = 39$) being responded to faster than invalid items ($M = 726, SE = 48$), and required response, $F_1(1, 32) = 9.52, p < .01, \eta_p^2 = .23, F_2(1, 46) = 4.42, p < .05, \eta_p^2$
=.09, with positive responses \((M = 685, SE = 42)\) given faster than negative responses \((M = 716, SE = 45)\).

More importantly, and as predicted, there was an interaction between validity and required response, \(F_1(1, 32) = 4.79, p < .05, \eta_p^2 = .13, F_2(1, 46) = 7.97, p < .01, \eta_p^2 = .15\). Planned comparisons revealed that this interaction was due to positive responses after valid sentences \((M = 625, SE = 30)\) being significantly faster compared to negative responses after valid sentences \((M = 724, SE = 52)\), \(F_1(1, 32) = 8.71, p < .01, \eta_p^2 = .21, F_2(1, 46) = 12.13, p < .01, \eta_p^2 = .21\), as well as compared to positive responses after invalid sentences \((M = 744, SE = 58)\), \(F_1(1, 32) = 9.08, p < .01, \eta_p^2 = .22, F_2(1, 46) = 18.96, p < .001, \eta_p^2 = .29\).

This pattern is in line with the predicted Stroop-like effect, with responses in the congruent condition valid sentence/positive response being faster than in the two incongruent conditions valid sentence/negative response and invalid sentence/positive response. However, it remains unclear whether this pattern reflects facilitation in the congruent condition, interference in incongruent conditions, or both. Given that there were no significant effects in the low knowledge condition – as indeed there should not be, if our participants did not have knowledge concerning the validity of the sentences – it appears that this condition may be a suitable neutral condition to test for interference and facilitation effects. Therefore, we reran the analyses comparing responses in the high knowledge conditions to responses in the respective low knowledge (control) conditions. For this purpose, we ran separate ANOVAs for valid and invalid sentences.

**Facilitation versus interference.** For valid sentences, there was a significant interaction between knowledge and required response, \(F_1(1, 32) = 5.87, p < .05, \eta_p^2 = .16, F_2(1, 46) = 5.78, p < .05, \eta_p^2 = .11\). This interaction was driven by a large facilitation effect for positive responses in the high knowledge condition \((M = 625, SE = 30)\) compared to the low knowledge control condition \((M = 706, SE = 48)\), \(F_1(1, 32) = 11.83, p < .01, \eta_p^2 = .27, F_2(1, 46) = 14.82, p < .001, \eta_p^2 = .24\), with a small numerical trend of interference for negative
responses in the high knowledge condition ($M = 724$, $SE = 52$) as compared to the low
knowledge control condition ($M = 716$, $SE = 42$) being nonsignificant, $F_1(1, 32) = 0.22$, $p =
.64$, $F_2(1, 46) = 0.59$, $p = .45$. In addition, there were main effects of knowledge, $F_1(1, 32) =
15.78$, $p < .001$, $\eta^2_p = .33$, $F_2(1, 46) = 10.18$, $p < .01$, $\eta^2_p = .18$, with responses being faster in
the high knowledge ($M = 675$, $SE = 39$) than in the low knowledge ($M = 711$, $SE = 44$)
condition, and of required response, $F_1(1, 32) = 6.14$, $p < .05$, $\eta^2_p = .16$, $F_2(1, 46) = 8.56$, $p <
.01$, $\eta^2_p = .16$, with positive responses ($M = 665$, $SE = 38$) being faster than negative responses
($M = 720$, $SE = 47$).

For invalid sentences, the numerical pattern indicates both interference for positive
responses (high knowledge: $M = 744$, $SE = 58$, low knowledge: $M = 707$, $SE = 45$) and
facilitation for negative responses (high knowledge: $M = 709$, $SE = 41$, low knowledge: $M =
750$, $SE = 55$); however, the interaction between knowledge and required response fell short
of significance, $F_1(1, 32) = 2.05$, $p = .16$, $F_2(1, 46) = 1.44$, $p = .24$.

**Error rates**

The full data for the error rates are displayed in Figure 3. The error rates were low
overall ($M = .051$, $SD = .108$); ANOVAs were performed on arc-sine transformed proportions.
The only significant effect that was reliable in both the $F_1$ and $F_2$ analyses was an interaction
of knowledge and validity, $F_1(1, 32) = 5.18$, $p < .05$, $\eta^2_p = .14$, $F_2(1, 92) = 7.54$, $p < .01$, $\eta^2_p =
.08$. This interaction was due to a significant difference between the error rates for valid and
invalid sentences in the high knowledge condition: There was a much higher error rate in the
valid ($M = .062$, $SE = .014$) than in the invalid condition ($M = .042$, $SE = .017$), $F_1(1, 32) =
6.02$, $p < .05$, $\eta^2_p = .16$, $F_2(1, 92) = 6.42$, $p < .05$, $\eta^2_p = .07$.

In addition, there was a two-way interaction between validity and required response in
the by-items analysis, $F_1(1, 32) = 1.55$, $p = .22$, $F_2(1, 92) = 4.99$, $p < .05$, $\eta^2_p = .05$, which was
qualified by a three-way interaction of all variables, $F_1(1, 32) = 3.62$, $p = .07$, $\eta^2_p = .10$, $F_2(1,
92) = 11.32$, $p < .01$, $\eta^2_p = .11$. 
**High versus low knowledge.** We followed up on this result by running separate by-items analyses for the high and low knowledge conditions. Similar to the results for the response latencies, there was a significant interaction of validity and required response only in the high knowledge condition, \( F_2(1, 46) = 17.29, p < .001, \eta^2_p = .27 \), but not in the low knowledge condition, \( F_2(1, 46) < 1 \). The interaction in the high knowledge condition was driven by the fact that after invalid sentences, more errors were made when the required response was positive (\( M = .063, SE = .030 \)) than when it was negative (\( M = .020, SE = .008 \)), \( F_2(1, 46) = 5.14, p < .05, \eta^2_p = .10 \), and after valid sentences, more errors were made when the required response was negative (\( M = .091, SE = .027 \)) than when it was positive (\( M = .033, SE = .008 \)), \( F_2(1, 46) = 13.05, p < .01, \eta^2_p = .22 \).

**Facilitation versus interference.** To address the question of facilitation vs. interference, as for the response latencies, we ran separate by-items ANOVAs for valid and invalid sentences.

For invalid sentences, the interaction of knowledge and required response was significant, \( F_2(1, 46) = 6.13, p < .05, \eta^2_p = .12 \). Participants made significantly less errors when a negative response was required and knowledge was high (\( M = .020, SE = .008 \)) compared to the low knowledge control condition (\( M = .068, SE = .020 \)), \( F_2(1, 46) = 10.42, p < .01, \eta^2_p = .19 \). The trend for a higher error rate for positive responses when knowledge was high (\( M = .063, SE = .030 \)) compared to the low knowledge control condition (\( M = .043, SE = .015 \)) was nonsignificant, \( F_2(1, 46) < 1 \).

For valid sentences, the interaction between knowledge and required response was also significant, \( F_2(1, 46) = 5.19, p < .05, \eta^2_p = .10 \), and it was again the negative responses that were affected by a compatibility effect: Participants made significantly more errors when the required response was negative and knowledge was high (\( M = .091, SE = .027 \)) compared to the low knowledge control condition (\( M = .048, SE = .015 \)), \( F_2(1, 46) = 10.80, p < .01, \eta^2_p = .19 \).
Discussion

The fact that a compatibility effect in terms of a validity x probe interaction obtained in a nonevaluative task in our study suggests that validation can also occur without an evaluative mindset, that is, without a task that explicitly encourages evaluation. Thus, in contrast to the conclusions by Wiswede et al. (2012), our study suggests that an evaluative task is not a prerequisite for validation and that merely understanding a sentence which can be judged as obviously true or false based on easily accessible knowledge is sufficient to produce a compatibility effect. It was already suggested by Singer (2006) that memory-based processes afford the verification of incoming information. The present study produced evidence that this verification, which operates on the information activated by the passive retrieval processes during comprehension, is nonstrategic, meaning that it operates without the reader's intention. As such, validation itself appears to be a passive process and by default a routine component of comprehension, in the sense that comprehension cannot occur independently from validation as both processes rely on the same knowledge activated by memory-based processes.

However, our results do not rule out the possibility that validation may be conditional in other ways. Quite to the contrary, it appears that a condition which must be fulfilled for validation to occur (or to be successful) is a certain depth of processing, i.e., a minimum level of comprehension. Shallow processing results in the activation of less information via memory-based processing which, in turn, results in less information on which validation can operate. When comprehension is impaired, for example because the experimental task requires only a relatively shallow level of processing (as seems to have been the case in the control group in the study by Wiswede et al.), then validation appears to be impaired as well. Thus, in contrast to what Wiswede et al. proposed, the conditionality of validation does not seem to refer to an evaluative goal or mindset, but rather to a certain level of processing, which points towards a close relationship between comprehension and validation.
One might object that the presentation of the "True" and "False" probes or the use of obviously true and false sentences may have been sufficient to induce an evaluative mindset by making the validity dimension salient to the reader. However, the fact that the same probe task and similar stimuli did not induce compatibility effects in the nonevaluative mindset group of the study by Wiswede et al. (2012) speaks against this notion.

In extension of previous studies, the present study also sheds light on the question of whether the Stroop-like validity/response compatibility effects reported by Richter et al. (2009), Wiswede et al. (2012), and Isberner and Richter (2013) are attributable to facilitation for congruent conditions, interference for incongruent conditions, or a combination of both. While the present study only produced clear evidence for facilitation of positive responses after valid sentences in the response latencies, the overall pattern of results for the response latencies and error rates suggests both facilitation of congruent and interference with incongruent responses. However, as there were no other interference or facilitation effects that were reliable in both the by-subjects and the by-items analyses in the present experiment, further research on this issue seems desirable.

One major difference between the experiment by Wiswede et al. (2012) and our study should be noted. To avoid interference with the EEG recordings, Wiswede et al. presented the probe 1800 ms after the onset of the final word, whereas it was presented only 300 ms after the onset of the final word in our experiment. This makes the behavioral data somewhat difficult to compare, as it is possible that compatibility effects produced by routine validation change over such a long time course. Nonetheless, the fact that a compatibility effect obtained in Wiswede et al.’s evaluative mindset condition despite the relatively long post-sentence delay suggests that the effects of validation in language comprehension are stable over quite some time. Alternatively, it is possible that the only reason why the “true” and “false” evaluations were kept active over such a long time is because they were relevant for one of the two tasks in the “evaluative mindset group,” namely the truth evaluation task. If this
assumption is correct, then compatibility effects should disappear at some point between 300 ms and 1800 ms after the sentence if the evaluation is irrelevant for the task (as in our experiment). An interesting direction for future research thus seems to be systematically investigating the time course of the effects by varying the stimulus onset asynchrony (SOA) of the probe. It is also possible that the pattern of interference and facilitation found in the present study merely represents a snapshot and would vary over different SOAs, which would allow insight into the time course of the positive and negative evaluations that arise from epistemic monitoring.

Overall, our results provide strong support for the idea that language comprehension entails a routine, nonstrategic validation process (epistemic monitoring; Richter et al., 2009), because readers do not seem to be able to ignore validity when they have easily accessible knowledge, even when assessing validity is irrelevant or even detrimental to the experimental task. This speaks against a conceptualization of comprehension and validation as nonoverlapping stages of information processing (e.g., Connell & Keane, 2006; Gilbert, 1991; Gilbert et al, 1990, 1993; Herbert & Kübler, 2011; Wiswede et al., 2012). However, it may seem at odds with studies showing readers’ susceptibility to false information (e.g., Bottoms, Eslick, & Marsh, 2010; Fazio, Barber, Rajaram, Ornstein, & Marsh, in press; Fazio & Marsh, 2008; Marsh & Fazio, 2006; Marsh et al., 2003; Rapp, 2008), as well as examples of readers’ failures to sometimes notice even blatant inconsistencies with their knowledge (e.g., Barton & Sanford, 1993; Erickson & Mattson, 1981). Our study may point towards a potential way of reconciling these seemingly contradictory findings: As validation seems to hinge on a minimum depth of comprehension, it is possible that such failures of validation are due to the construction of an underspecified mental representation (Bohan & Sanford, 2008; Sanford, 2002; Sanford & Graesser, 2006; Sanford, Leuthold, Bohan, & Sanford, 2011). In line with this idea, it has been shown that factors which influence the specification of a mental
representation, such as linguistic focus (e.g., Sanford & Garrod, 2005), also influence the extent to which false information is detected (Bredart & Modolo, 1988).

Moreover, despite the fact that an evaluative processing goal is not necessary for validation to occur, our results are still compatible with the idea that validation is conditional in other ways. For example, it may be affected by the text genre (narrative vs. expository texts) or by the perceived credibility of a text source (encoding under distrust; e.g., Schul, Mayo, & Burnstein, 2004). For example, people seem to be particularly susceptible to false information and persuasion when reading or viewing narratives (e.g., Appel & Richter, 2007; Gerrig & Prentice, 1991; Green & Brock, 2000; Umanath, Butler, & Marsh, 2012), which suggests that epistemic monitoring might be suppressed to some extent in narrative (as opposed to argumentative) texts. Future research should focus more explicitly on the conditions under which validation succeeds or fails, with the goal of reconciling evidence for readers’ apparent susceptibility to false information with the abundant evidence for routine validation in language comprehension.
References


Figure 1. Trial structure of the experiment. (A) Experimental trials. (B) Filler trials with probe after the second word. (C) Filler trials with probe after the first word.
Figure 2. Mean correct response latency as a function of validity (invalid, valid) and required response (positive, negative) for a) low knowledge and b) high knowledge. Error bars correspond to ±1 standard error of the mean computed for within-subjects designs (Morey, 2008).
Figure 3. Mean error rates as a function of validity (invalid, valid) and required response (positive, negative) for a) low knowledge and b) high knowledge. Error bars correspond to ±1 standard error of the mean computed for within-subjects designs (Morey, 2008).