Can readers ignore implausibility? Evidence for nonstrategic monitoring of event-based plausibility in language comprehension

Maj-Britt Isberner and Tobias Richter

University of Kassel

Manuscript accepted for publication in the journal *Acta Psychologica*

Word Count: 7,274

Corresponding author:

Maj-Britt Isberner

University of Kassel, Department of Psychology

Holländische Str. 36-38

34127 Kassel, Germany

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

Telephone: +49-561 804 1905

Fax: +49-561 804 3586

Abstract

We present evidence for a nonstrategic monitoring of event-based plausibility during language comprehension by showing that readers cannot ignore the implausibility of information even if it is detrimental to the task at hand. In two experiments using a Stroop-like paradigm, participants were required to provide positive and negative responses independent of plausibility in an orthographical task (Experiment 1) or a nonlinguistic color judgment task (Experiment 2) to target words that were either plausible or implausible in their context. We expected a nonstrategic assessment of plausibility to interfere with positive responses to implausible words. ANOVAs and linear mixed models analyses of the response latencies revealed a significant interaction of plausibility and required response that supported this prediction in both experiments, despite the use of two very different tasks. Moreover, it could be shown that the effect was not driven by the differential predictability of plausible and implausible words. These results suggest that plausibility monitoring is an inherent component of information processing.

Keywords: language comprehension; plausibility monitoring; event knowledge; predictability; verification; validation; context

Classification codes: 2340, 2720

1. Introduction

Whether world or event knowledge is immediately accessed during language comprehension is still a point of contention. While some studies report immediate effects of such knowledge on various measures of reading comprehension including reading times, eye tracking measures, and event related potentials (ERPs) (e.g., Hagoort et al., 2004; McRae et al., 1998; Matsuki et al., 2011; Rapp, 2008; Van Berkum et al., 2005), other studies come to the conclusion that its influence in language comprehension is delayed in comparison to semantic knowledge. For example, Rayner et al. (2004) and Warren and McConnell (2007) found early effects of semantic violations on eye movements, but not of implausibility, suggesting that semantic knowledge is temporally privileged in language comprehension, whereas the access of world knowledge is slightly delayed.

A recent study by Matsuki et al. (2011) attempted to reconcile these seemingly contradictory findings by scrutinizing the typicality of the events described in the stimuli used in different studies. Their hypothesis was that typicality might be the key to explain the differences in the obtained results: In order to obtain early plausibility effects in reading times and eye tracking measures, they proposed that it is crucial that the plausible stimuli describe situations which are typical of people's world experience. The authors ensured typicality of their own stimuli by using production norms in addition to rating norms. Since the focus of their study was on instrument-action combinations, they asked their participants to "List the things or people that have the following actions done to them with the specified instruments" (p. 916). Based on the responses, they created minimal pairs of stimuli which reflected typical and atypical (but not anomalous) real world events, such as *Donna used the hose/shampoo to wash her filthy car/hair* (typical) and *Donna used the shampoo/hose to wash her filthy car/hair* (atypical). With these stimuli, in contrast to Rayner et al. (2004) and Warren and McConnell (2007), they found rapid effects of event-based plausibility (or typicality) in both self-paced reading and eye tracking, suggesting that there is in fact no delay in the access of event knowledge when this knowledge is typical of the readers' experience.

Similarly, Staub et al. (2007) report immediate effects of plausibility in an ingenious study that used sentences which were always globally plausible, but contained noun-noun compounds (e.g., *cafeteria manager*) whose modifier was either plausible or implausible in its context if it was initially analysed as a head noun (e.g., *The new principal visited / talked to the cafeteria manager*). Plausibility had very rapid effects on eye movements, much faster than effects usually found in ERP studies, with implausibility resulting in an increase in reading time on the initially implausible word. Moreover, the size of the reading time penalty correlated with offline ratings of the implausibility of the word in the context leading up to it. It is important to note that these effects obtained although readers were merely asked to read for comprehension and although all sentences were globally plausible. Staub et al. (2007) interpreted their findings as evidence that the rapid effects of plausibility were not due to strategic factors.

These results are fascinating for two reasons: First, because they suggest that plausibility is monitored in the absence of an explicit evaluative processing goal. Second, this monitoring seems to follow the same principles as intentional plausibility ratings, suggesting that the plausibility assessment that can be computed in an intentional decision process is in fact – in some form – immediately available as a word is comprehended in its context. These findings are in stark contrast with two-step models of sentence verification which assume that any kind of evaluation is delayed with regard to comprehension, or in other words, withheld until the comprehension process has terminated (e.g., Gilbert et al., 1993). Rather, they suggest that language comprehension comprises a routine, online plausibility monitoring process that operates nonstrategically and fast on a word-by-word basis as the linguistic input unfolds.

Interestingly, Staub et al. (2007) do not draw a distinction between implausible and semantically anomalous sentences, which is in line with Matsuki et al.'s (2011) conclusion that this kind of distinction may in fact be arbitrary (see also Jackendoff, 2002, and Hagoort et al., 2004). However, if one inspects their stimuli, at least some of the local implausibilities are due to animacy violations (including the aforementioned example), which are generally considered semantic violations. It thus remains an open question whether the same nonstrategic process underlies the plausibility effects that were obtained by Matsuki et al. (2011).

In the present study, our goal is to investigate this question by testing whether event-based plausibility is routinely monitored during language comprehension. We do this by testing the nonstrategic nature of the proposed routine plausibility monitoring process with a Stroop-like paradigm (Stroop, 1935) in which an assessment of plausibility is irrelevant or even detrimental to task performance. Specifically, we test the potential interference of plausibility monitoring with an unrelated task that requires responses which are orthogonal to plausibility. Beyond the question of the time course of plausibility effects, we thereby attempt to elucidate what actually happens in the reader when he or she encounters implausible information (besides taking longer to process it than plausible information).

What kind of interference is to be expected from a routine, nonstrategic monitoring of event-based plausibility? We assume that, if readers indeed routinely assess plausibility, they will react to information that is inconsistent with their event knowledge with a negative response tendency. This negative response tendency, in turn, should make it more difficult to provide positive responses of any kind, even if the responses are completely unrelated to plausibility. To test this hypothesis, we make use of the so-called epistemic Stroop paradigm, an adaptation of the Stroop paradigm introduced by Richter et al. (2009) for testing the interference of factual knowledge with an unrelated judgment task. In their study, participants were asked to judge the orthographical correctness of words embedded in assertions that were presented word by word on a computer screen and were either valid or invalid with regard to common factual knowledge. In experimental items, the word that had to be judged was the last word of the assertion, and it was spelled either correctly or incorrectly (e.g., *Perfume contains scents / sents* or *Soft soap is edible / eddible*; the original sentences were in German: *Parfüm enthält Duftstoffe/duftstoffe* and *Schmierseife ist essbar/essbahr*). Although the validity of the assertions was irrelevant to the orthographical task, responses were delayed when the word to be judged was presented at the end of an invalid assertion but required a positive ("correct") response. This resulted in a significant interaction of validity and orthographical correctness.

If our assumption of a nonstrategic plausibility monitoring process holds, we should find a similar effect for stimuli which tap into readers' event knowledge. Specifically, we expect to find slower latencies for positive (i.e., affirmative) responses in the unrelated task when a word (for example, the word *plumber)* is implausible in its context (*Frank has a broken leg. He calls the plumber.)* compared to when it is plausible (*Frank has a broken pipe. He calls the plumber.*). We therefore expect a significant interaction of plausibility and required response which conforms to this pattern.

In Experiment 1, we test this hypothesis with the same orthographical judgment task used by Richter et al. (2009). However, if it is true that the interference of plausibility monitoring hinges on the positive/negative character of the response rather than on other task characteristics, it should obtain in any kind of task that requires positive and negative responses and is independent of plausibility. In order to test this hypothesis, we go one step further in Experiment 2 and investigate the interference of plausibility monitoring with a completely different, nonlinguistic task which is even more obviously independent of plausibility than the orthographical task: The task of judging whether or not a word that is plausible or implausible in its context has changed color.

As discussed by Matsuki et al. (2011), a variable that is often confounded with plausibility is predictability. Although, as the authors point out, these two dimensions are practically extremely difficult to disentangle, we nonetheless attempt to do this by varying the predictability of the target word in the plausible context while keeping plausibility constant. According to Matsuki et al. (2011), “One way to differentiate the two would be to contrast implausible items with plausible ones for which cloze values of all targets is zero” (p. 926). However, since it is, as the authors state, “virtually impossible” (p. 926) to construct plausible targets with a cloze value of zero (particularly in minimal pairs of stimuli that differ only regarding the target word), since even atypical or implausible targets usually have cloze values higher than that, our goal was to approximate zero as much as possible without creating unnatural stimuli.

Moreover, in order to keep the plausible and implausible conditions strictly parallel, we designed our stimulus material in such a way that the same target sentences could be used in both conditions. This was achieved by varying the plausibility of each target sentence by means of a context sentence, which rendered the same target sentence either plausible or implausible. In this way, and in extension of the aforementioned studies, our experiments also allowed testing whether the extrasentential linguistic context routinely becomes part of the background against which incoming information is monitored for plausibility.

2. Experiment 1

The primary goal of Experiment 1 was to investigate whether event-based plausibility is nonstrategically monitored by testing its interference with an orthographical task unrelated to plausibility, using a Stroop-like paradigm introduced by Richter et al. (2009). We assume that if this is the case, information that is implausible with regard to a comprehender’s event knowledge should elicit a negative response tendency. The negative response tendency, in turn, should interfere with positive responses in the unrelated task. Thus, we expect participants to take longer to indicate that a word is spelled correctly when it is implausible in its context than when it is plausible.

2.1 Method

2.1.1 Participants

Participants were 70 psychology undergraduates at the University of Cologne (52 women and 18 men). All participants were native speakers of German. Their average age was 24.2 years (*SD* = 4.8).

2.1.2 Stimulus material

Stimuli were pairs of context and target sentences describing situations that were either plausible or implausible with regard to common event knowledge. For each of the experimental items, four different versions were constructed. First, there were two versions of each context sentence. One version rendered the last word of the target sentence plausible and the other one rendered it implausible (e.g., *Frank has a broken pipe / leg. He calls the plumber.*). Second, there were two versions of each target sentence. One version ended with a word that was assumed to have a high predictability in the plausible context, and the other ended with a word that was equally plausible but had a low predictability in the plausible context (e.g., *Frank has a broken pipe. He calls the plumber / tradesman.*). Of each of the four versions of each item, an orthographically incorrect version was constructed by inserting, exchanging, or removing one letter or changing the case of the last word of the target sentence, while maintaining the phonology of the correct word (such as *shammpoo* instead of *shampoo*; the actual stimuli were in German, e.g., *Shammpoo* instead of *Shampoo*). In addition to the experimental items, 160 filler items were constructed. These were also pairs of context and target sentences, of which 80 described plausible situations and 80 described implausible situations. Of the plausible as well as the implausible filler items, half contained a word with a spelling mistake. This word served as the target word for the orthographical task. The procedure for inserting the spelling mistakes was based on the same principles as in the experimental items. The position of the misspelled word within each filler item was selected randomly, excluding the first word of the context sentence and the last word of the target sentence. Following the same principle, one word was selected as the target word in each of the 80 remaining filler items but maintained in its orthographically correct form.

2.1.3 Norming study

A norming study was conducted to select experimental items with both an effective plausibility and an effective predictability manipulation out of a pool of 97 items. The participants of the norming study (14 psychology undergraduates not identical to the experimental sample) completed a questionnaire with two tasks. First, there was a cloze test to assess the predictability of the final word in each item. Participants were asked to read each item and spontaneously fill in the last word of the target sentence, which had been substituted by a blank. Second, they were asked to rate the plausibility of each of the four (orthographically correct) sentence pairs that resulted from pairing both versions of the context sentence with both versions of the target sentence (4 x 97 = 388 sentence pairs). Participants were asked to indicate for each sentence pair whether they found it plausible (“yes”) or implausible (“no”). The sentence pairs were presented in the same order to all participants but mixed randomly within the questionnaire. Based on these data, 64 out of the 97 items were selected in which both the plausibility manipulation as well as the predictability manipulation proved to be effective. These were items in which the mean agreement with the assumed plausibility was high for all versions of the item and in which the cloze values were high only for the predictable word in the plausible condition and low in all other conditions. The norms for the selected items are displayed in Table 1.

2.1.4 Procedure

All items were presented word by word on a computer screen using Rapid Serial Visual Presentation (RSVP) with a fixed rate of 600 ms per word. Each word was presented in bold black letters in the font type Arial (approximate height 1 cm) in a white 13 x 6 cm square placed in the middle of the screen against a silver background. The viewing distance was approximately 60 cm. Each trial was preceded by a fixation cross presented for 250 ms and followed by a blank screen presented for 500 ms. At one word per trial (the target word), the presentation stopped and participants were prompted by the question *Spelling?*, which appeared above the target word 300 ms after the onset of the target word, to indicate whether or not the word was spelled correctly. The prompt and the target word remained on the screen until the participant provided a response. Participants were instructed to provide their responses as fast and as accurately as possible by pressing ‘k’ for correct spelling and ‘d’ for incorrect spelling, and to keep their fingers on the two response keys throughout the whole experiment. As a reminder for which of the two keys to press for which response, the prompt was accompanied by a label *correct* in green font inside a white box with a green frame on the right hand side, and a label *incorrect* in red font inside a white box with a red frame on the left hand side. On half of the trials, the target word was spelled correctly, requiring a “correct”-response, and on the other half of the trials, the presented word was spelled incorrectly, requiring an “incorrect”-response. In experimental trials, the target word was always the final word of the item. In filler trials, the target word was at a randomly selected position within the item (see *2.1.2 Stimulus Material*). The purpose of the filler items was to ensure that participants would not be able to guess at which word of the item they would be asked to provide a response. To encourage correct responses, participants received a feedback on the accuracy of each of their responses, which was presented for 600 ms after each response. The trial either ended with a blank screen (experimental items) or continued with the next word of the item (filler items). The first six items presented to each participant were practice items that were not included in the analysis.

2.1.5 Design

The design was a 2(*plausibility*: plausible vs. implausible) X 2(*predictability*: predictable vs. unpredictable) X 2(*required response*: positive vs. negative) within-subjects design. Dependent variables were the response latency and the accuracy of the responses. Assignments of experimental items to experimental conditions were counterbalanced across participants by eight item lists. Each participant saw eight experimental items in each of the eight experimental conditions. Experimental and filler items were presented in random order.

2.2 Results and Discussion

Type-I error probability was set at .05 for all hypothesis tests. Under the assumption of a medium effect size ($f = .25$ according to Cohen, 1988) and medium correlations ($\rho = .5$) between the levels of the independent variables in the population, the design and sample size of Experiment 1 yielded a power (1-β) of .98 for detecting the focal interaction of plausibility and required response in the ANOVA based on subjects as the units of analysis (power computed with the software G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007). We conducted ANOVAS 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 are based on subjects as the units of analysis. Standard errors of the mean were computed for within-subjects designs (Morey, 2008).

In addition to the ANOVA analyses, we conducted a linear mixed models (LMM) analysis for the response latencies and a generalized linear mixed models (GLMM) analysis with logit link for the error rates with subjects and items included as random factors, i.e. the means of subjects as well as items were allowed to vary randomly. This type of analysis accounts for the fact that both subjects and items represent samples of larger populations. Unlike the F1- and F2-ANOVA, the LMM/GLMM analysis with crossed random effects for subjects and items does not decrease power but allows for an adequate and stringent test of the hypothesized effects of the independent variables in one single model (for further discussion, see Baayen et al., 2008). We included all three independent variables as contrast-coded predictors with fixed effects in the model (plausibility: 1 = plausible, -1 = implausible; predictability: 1 = predictable, -1 = unpredictable; required response: 1 = positive, -1 =

negative). In addition, the presentation position of each item was included in the model as centered predictor (fixed effect) to control for position effects. The LMM/GLMM analysis was conducted with the *lmer* command of the *lme4* package for R (Bates et al., 2011). For the sake of conciseness, only significance tests associated with the fixed effects (main and interaction effects) of the independent variables are reported as these are directly relevant for our hypotheses (data files and R-scripts for both experiments are available from the authors upon request). Please note that no degrees of freedom are reported for the t-values of the LMM analysis because it is still unclear how these should be derived. However, given the large number of observations in the present experiments (items times participants), it is safe to assume that the distribution of t-values approximates the standard normal distribution (z-distribution; see Bayen et al., 2008, Note 1). Thus, the standard normal distribution was assumed for significance tests of fixed effects in the LMM analysis.

2.2.1 Response latencies

Response latencies were calculated for correct responses (93% of the responses in experimental trials). Response latencies deviating more than three standard deviations from either the subject or item mean (1.8% of all correct latencies) were treated as outliers and removed from the data set. Figure 1 shows the mean correct response latencies as a function of plausibility and required response; Table 2 displays the means and standard deviations associated with the by-subjects analysis. We found significant main effects for all of the three independent variables. Plausible target words (M = 962 ms, SE = 6 ms) elicited faster responses than implausible target words (M = 1035 ms, SE = 6 ms), $F_1(1, 69) = 48.30, p < .001, \eta_p^2 = .41, F_2(1, 63) = 24.41, p < .001, \eta_p^2 = .28$ (LMM analysis: $t = -6.20, p < .05$). Predictable words (M = 947, SE = 8) elicited faster responses than non-predictable words (M = 1050 ms, SE = 8 ms), $F_1(1, 69) = 53.58, p < .001, \eta_p^2 = .44, F_2(1, 63) = 21.26, p < .001, \eta_p^2 = .25$ (LMM analysis: $t = -8.88, p < .05$). Furthermore, negative responses to incorrectly spelled words (M = 951 ms, SE = 11 ms) were faster than positive responses to correctly spelled words (*M* = 1046 ms, *SE* = 11 ms), $F_1(1, 69) = 22.50$, $p < .001$, $\eta_p^2 = .25$, $F_2(1, 63) = 12.80$, $p < .01$, $\eta_p^2 = .17$, (LMM analysis: $t = 7.75$*,* $p < .05$).

However, the main effects of plausibility and required response were qualified by a significant interaction of the two variables, $F_1(1, 69) = 6.77$, $p < .05$, $\eta_p^2 = .09$, $F_2(1, 63) = 5.69$, $p < .05$, $\eta_p^2 = .08$ (LMM analysis: $t = -2.97$*,* $p < .05$). Planned contrasts revealed that the pattern underlying the interaction conformed to the hypothesized Stroop-like effect. Positive responses in the orthographical task were significantly slower for implausible (*M* = 1103 ms, *SE* = 18 ms) compared to plausible words (*M* = 990 ms, *SE* = 14 ms), $F_1(1, 69) = 31.62$, $p < .001$, $\eta_p^2 = .31$, $F_2(1, 63) = 30.61$, $p < .001$, $\eta_p^2 = .33$. Negative responses were also slower for implausible (*M* = 968 ms, *SE* = 16 ms) compared to plausible words (*M* = 934 ms, *SE* = 14 ms), but with $F_1(1, 69) = 4.13$, $p < .05$, $\eta_p^2 = .06$, $F_2(1, 63) = 3.51$, $p = .07$, $\eta_p^2 = .05$, this difference was much smaller than for positive responses and non-significant in the by-items analysis. Furthermore, there was no three-way interaction with predictability, $F_1(1, 69) = 2.35$, $p = .13$, $F_2(1, 63) < 1$, $p = .47$ (LMM analysis: $t = 1.82$*,* $p > .05$).

2.2.2 Error rates

The error rates were low overall (*M* = .07, *SD* = .11). There was a significant main effect of required response, $F_1(1, 69) = 32.92$, $p < .001$, $\eta_p^2 = .32$, $F_2(1, 63) = 19.14$, $p < .001$, $\eta_p^2 = .23$ (ANOVAs performed on arc-sine transformed proportions; GLMM analysis: $z = 7.42$*,* $p < .001$). More errors were made in the judgment of orthographically incorrect words, that is, when the required response was negative (*M* = .097, *SE* = .006) compared to orthographically correct words, that is, when the required response was positive (*M* = .043, *SE* = .006). Furthermore, there was a significant main effect of predictability in the by-subjects ANOVA, $F_1(1, 69) = 6.93$, $p = .01$, $\eta^2_p = .09$, $F_2(1, 63) = 3.29$, $p = .08$, $\eta^2_p = .05$ (GLMM analysis: $z = 4.03$*,* $p < .001$). More errors were made in the judgment of non-predictable words (*M* = .082, *SE* = .004) compared to predictable words (*M* = .058, *SE* =.004). In contrast to the results for the response latencies, there was no interaction effect of plausibility and required response, $F_1(1, 69) < 1, p = .88, F_2(1, 63) < 1, p = .35$ (GLMM analysis: $z = 0.28, p = .78$). Thus, there was no indication of a speed-accuracy trade-off in our data.

The delay of positive responses to words that are implausible in their context supports the hypothesis that event-based plausibility is routinely monitored during language comprehension and results in the detection and rejection of implausible information. However, the fact that both positive and negative responses were faster when the target word was plausible compared to when it was implausible prevents a fully conclusive interpretation of this effect. This pattern indicates that the orthographical task was easier for plausible than for implausible words, which may be due to the fact that words are generally easier to recognize when they are congruent with a context than when they are incongruent (e.g., Stanovich & West, 1981, 1983). Thus, plausible words may have been easier to recognize and check for orthographical correctness. However, this makes the orthographical task somewhat suboptimal for investigating effects of nonstrategic plausibility monitoring because it might attenuate the expected difference between the effects of plausibility on positive and negative responses. We ran Experiment 2 to clarify this issue.

It may also seem unusual that there was no advantage for affirmative responses in our task, which is often found in other types of tasks, such as lexical decision. In fact, negative responses in our task were significantly faster than positive responses. This main effect of required response in the orthographical task was also found by Richter et al. (2009, Experiment 3). It may be attributable to the fact that the misspelled words were phonologically and orthographically very similar to the original words so that they remained easy to recognize, while the orthographical errors were blatant enough to be easy to spot for native speakers with a regular school education. This interpretation is supported by the high accuracy rate despite the speeded response conditions (93%). It is also important to note that our task was quite different from lexical decision with regard to both the stimuli and the instructions. Most importantly, there were no nonwords in our task, unless one would like to define the misspelled words as nonwords. Even so, the instruction for the orthographical task would have led participants to perceive them as real but misspelled words rather than as meaningless nonwords (such as those that are usually used in lexical decision). Thus, the processing induced by our task instruction and stimuli should have been rather different from the processing required by a lexical decision task.

Importantly, the main effect of required response does not limit the interpretation of the results because the critical comparisons were those between the two plausibility conditions for the same response type, rather than between positive and negative responses. However, because we were not interested in effects other than those produced by the assumed plausibility monitoring process, a task in which there is no general advantage for one response or the other would be preferable, and we tried to achieve this in Experiment 2.

3. Experiment 2

Experiment 2 was designed to eliminate potential problems of Experiment 1 by using a different kind of task. The fact that in Experiment 1, negative responses were also slower when the word was implausible in its context indicates that plausibility might have been confounded with task difficulty. Therefore, we will use a different task in Experiment 2 whose difficulty should be unaffected by plausibility. Moreover, for the purpose of testing the generalizability of the interference effect, it is advantageous to use a task which strongly differs from the orthographical task. Therefore, we chose the nonlinguistic task of judging whether or not the target word changes color.

Finally, despite the fact that the orthographical task did not require any semantic (let alone plausibility) judgment, the presentation rate of one word per 600 ms used in Experiment 1 might have provided participants with sufficient time to engage in some kind of strategic evaluation of the message prior to seeing the target word. For this reason, Experiment 2 used a presentation rate of one word per 300 ms which roughly corresponds to the average fixation duration during reading (Rayner, 1998). Thus, the presentation rate in Experiment 2 was sufficiently short to minimize any strategic processing during sentence reading besides the focal color judgment task.

3.1 Method

3.1.1 Participants

Participants were 67 undergraduates (native speakers of German) at the University of Kassel. The average age of the participants (44 women and 23 men) was 24.2 years (*SD* = 5.7).

3.1.2 Stimulus material

The orthographically correct versions of the experimental and filler items of Experiment 1 were used. The target words were the same as in Experiment 1 (i.e., the final word in experimental items and a randomly selected word in filler items) but they now either changed color or remained black when the response prompt appeared.

3.1.3 Procedure

The procedure was identical to the procedure used in Experiment 1, except for the following differences: First, the presentation time for each word in the RSVP and for the feedback was reduced to 300 ms. Second, 300 ms after the target word appeared, instead of the orthographical judgment participants were now prompted to indicate whether or not the word had changed color as the prompt appeared (50% of the trials required a yes response; in the other half of the trials, the word remained black). In the color change trials, colors were chosen randomly from a list of 9 colors which had been approved for readability on a white background.

3.1.4 Design

Design and dependent variables were the same as in Experiment 1.

3.2. Results and Discussion

Type-I error probability was set at .05 for all hypothesis tests. The design and sample size of Experiment 2 yielded a power of .98 for detecting the focal interaction of plausibility and required response (with $f = .25$ and $\rho = .5$) in a by-subjects ANOVA. Due to a programming error, the presentation of one of the 64 experimental items was faulty in one of the eight conditions. For this reason, this item was discarded from all further analyses. As in Experiment 1, 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. Standard errors of the mean were computed for within-subjects designs (Morey, 2008). In addition, the fixed effects from an LMM/GLMM analysis with crossed random effects of subjects and items (Baayen et al., 2008) are reported.

3.2.1 Response latencies

Response latencies were calculated for correct responses (96.8% 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 removed from the data set. Figure 2 shows the mean correct response latencies as a function of plausibility and required response; Table 3 displays the means and standard deviations based on subjects as the units of observation. As in Experiment 1, there was a main effect of plausibility which was significant in the by-subjects and the LMM analysis. Plausible target words ($M = 661$ ms, $SE = 4$ ms) were responded to faster than implausible target words ($M = 677$ ms, $SE = 4$ ms), $F_1(1, 66) = 6.09$, $p < .05$, $\eta_p^2 = .08$, $F_2(1, 62) = 3.59$, $p = .06$, $\eta_p^2 = .06$ (LMM analysis: $t = -2.25$, $p < .05$).

Moreover, the analysis revealed an interaction of plausibility and required response which was significant in the by-subjects analysis, $F_1(1, 66) = 5.18$, $p < .05$, $\eta_p^2 = .07$, but missed significance using items as a random source of variance, $F_2(1, 62) = 2.66$, $p = .11$, $\eta_p^2 = .04$. Most importantly, however, the interaction of plausibility and required response was significant in the LMM analysis which includes subjects as well as items as sources of random variance (t = -2.27, p < .05). In order to interpret the interaction, we conducted planned contrasts which revealed that the pattern underlying the interaction was similar to the pattern found in Experiment 1. As before, positive responses were slower for implausible (M = 685 ms, SE = 8 ms) compared to plausible words (M = 654 ms, SE = 7 ms), F_1(1, 66) = 9.45, p < .01, η_p^2 = .13, F_2(1, 62) = 5.32, p < .05, η_p^2 = .08. Crucially, and in contrast to Experiment 1, the latencies of negative responses to plausible (M = 667 ms, SE = 7 ms) and implausible target words (M = 670 ms, SE = 7 ms) did not differ significantly from each other, F_1(1, 66) < 1, p = .71, F_2(1, 62) < 1, p = .90. Moreover, there was again no three-way interaction with predictability, F_1(1, 66) < 1, p = .56, F_2(1, 62) < 1, p = .56 (LMM analysis: t = -0.51, p > .05).

3.2.2 Error rates

Again, the error rates were low overall (M = .03, SD = .07) and showed no indication of a speed-accuracy trade-off: The interaction of plausibility and required response was not significant F_1(1, 66) < 1, p = .98, F_2(1, 62) < 1, p = .87 (ANOVAs performed on arc-sine transformed proportions; GLMM analysis: z = 0.18, p = .86). All other effects were also non-significant, with all p-values exceeding .10, except for the interaction of plausibility and predictability, F_1(1, 66) = 9.18, p < .01, η_p^2 = .12, F_2(1, 62) = 10.10, p < .01, η_p^2 = .14 (GLMM analysis: z = 2.78, p < .01).This interaction was due to more errors being made in response to non-predictable words in the plausible condition (M = .047, SE = .007) compared to non-predictable words in the implausible condition (M = .025, SE = .005), F_1(1, 66) = 8.08, p < .01, η_p^2 = .11, F_2(1, 62) = 6.70, p < .05, η_p^2 = .10, as well as compared to predictable words in the plausible condition (M = .021, SE = .005), F_1(1, 66) = 9.49, p < .01, η_p^2 = .13, F_2 (1, 62) = 6.67, p < .05, η_p^2 = .10. As we had no hypotheses concerning this interaction, and it does not affect the interpretation of the response latency data, we simply point it out here without further interpretation.

These results are an important extension of Experiment 1. First, the similarity of the patterns in the two experiments, despite the fact that the tasks were entirely different (i.e., a linguistic orthographical task vs. a nonlinguistic color judgment task), is striking. This confirms our assumption that the only task dimension which produces the pattern is the requirement of positive and negative responses independent of plausibility. Second, and most importantly, the pattern that emerged in Experiment 2 clearly indicates that the effect hinges on a delay of positive responses to implausible words, since the negative responses were unaffected by plausibility. Third, the effect occured despite the fact that the presentation rate in Experiment 2 was much shorter than in Experiment 1, reducing the likelihood of strategic processing even further. Finally, there was no main effect of required response as in Experiment 1, which suggests that this effect was due to the specific demands of the orthographical task.

4. General Discussion

We assumed that the influence of event-based plausibility in comprehension, as found by Matsuki et al. (2011), reflects a routine plausibility monitoring process that is nonstrategic and inherent in language comprehension. In order to give this tacit process a “voice”, we tested the interference of its assumed negative outcome for implausible information with incongruent positive responses in an unrelated judgment task using a Stroop-like paradigm adapted from Richter et al. (2009). In Experiment 1, the task we used was an orthographical judgment task as in the original Richter et al. (2009) study. In Experiment 2, we used a nonlinguistic color judgment task and increased the presentation rate in order to rule out potential alternative explanations and test the generalizability of the results.

In line with our predictions, responses were delayed in both tasks when the task required a positive response to a target word that was implausible in its context, compared to when it was plausible, resulting in an interaction of plausibility and required response.

However, in the orthographical task, negative responses were also slower for implausible compared to plausible words, suggesting a higher overall task difficulty for implausible words. This may have been due to the fact that words are generally easier to recognize when they are plausible in their context (e.g., Stanovich & West, 1981, 1983), which makes the orthographical task somewhat suboptimal for investigating the effects of nonstrategic plausibility monitoring. For this reason, and to test the generalizability of our results, we chose a nonlinguistic color judgment task in Experiment 2. In spite of the entirely different nature of the task, the global pattern of results was strikingly similar. Although the interaction of plausibility and required response fell short of significance in the by-items analysis, it was significant in the by-subjects analysis as well as in a Linear Mixed Models analysis which takes both subjects and items into account as sources of random variation. Moreover, the critical planned contrasts produced the same results in the F1 and F2 analyses, with positive responses being slower for implausible compared to plausible words and – in contrast to Experiment 1 – negative responses being unaffected by plausibility. This confirms that we indeed managed to find a task whose difficulty does not vary with plausibility and thus in principle allows the interference effect to emerge even more clearly. In addition, this task also eliminated the response time advantage for negative responses, which seemed to be specific to the orthographical task. Overall, the two experiments provide strong evidence for routine, nonstrategic plausibility monitoring during language comprehension.

These findings are in line with both the Matsuki et al. (2011) findings that event-based plausibility is immediately accessed in language comprehension, as well as with the Staub et al. (2007) findings that plausibility effects on language comprehension are nonstrategic. In addition, our results bridge both findings by suggesting that, despite differences between the stimuli, the same nonstrategic process may be underlying the rapid plausibility effects obtained in both studies. Beyond questions of the time course of access to different kinds of knowledge, our results suggest that event knowledge and the assessment of plausibility based on this knowledge are routine and obligatory in language comprehension. An interesting extension of the Matsuki et al. (2011) and the Staub et al. (2007) results is that while in those studies, plausibility of the target word hinged on the intrasentential context, in our study it was manipulated by the extrasentential context (i.e., the preceding context sentence). The fact that the effect obtained nonetheless is in line with other findings that people immediately relate linguistic input to the widest available context (e.g., Hagoort & Van Berkum, 2007; Just & Carpenter, 1980; Nieuwland & Van Berkum, 2006; Van Berkum et al., 1999; Van Berkum et al., 2003).

Furthermore, we attempted to rule out the alternative explanation that the effect might be driven by the predictability rather than the plausibility of the target word by using target words that were similar in plausibility but with highly different cloze values. Naturally, the non-predictable target words were still more predictable in the plausible than in the implausible condition; however, if the Stroop-like effect was driven by predictability (i.e., by a negative response tendency elicited by unexpected words) it would be expected to be much stronger for the predictable words. Alternatively, an interaction of predictability and required response analogous to the predicted interaction of plausibility and required response should emerge if predictability was indeed the crucial variable here. Contrary to this idea, neither of the experiments showed a modulation of the effect by predictability in terms of a three-way interaction or an interaction of predictability and required response. Hence, it seems unlikely that the effect obtained in our study is due to predictability differences between plausible and implausible items.

Despite the fact that the overall interaction of plausibility and required response and the corresponding planned comparisons are in line with our predictions, it must be noted that the interaction effect was slightly smaller in Experiment 2 than in Experiment 1 (which is evident in the by-items analysis). This pattern may point towards a disadvantage of the nonlinguistic color judgment task: it did not require comprehension of the stimuli and may thus have reduced semantic processing. It is important to note that while we argue that plausibility assessment is nonstrategic, we do not argue that it can occur without an adequate level of comprehension. Despite the proposed nonstrategic nature of plausibility monitoring, it is still reasonable to assume that more shallow semantic processing will reduce validation processes and hence their interference with other tasks. A way to avoid this problem and ensure deeper semantic processing while still using a nonlinguistic task would be to include questions which require comprehension but not plausibility assessment of the sentences. This would also open up the possibility of directly exploring the relationship between depth of semantic processing and nonstrategic plausibility assessment, which our results suggest to be a promising endeavor for future experiments.

One further issue worth noting is the asymmetry of the effects obtained for positive and negative responses. In our hypotheses, we predicted the interference of a negative response tendency evoked by implausible information with positive responses. We did not expect a converse interference of plausible information with negative responses because we assumed the monitoring process to respond negatively to implausible information (in terms of an error detection process) rather than positively to plausible information. Nonetheless, one might have expected facilitation for negative responses after implausible information, which is clearly not present in either of the experiments. A possible interpretation of this result is that it might point towards a special status of implausible information: It could be that readers react to implausible information with reduced acceptance rather than with outright rejection because they cannot be certain whether the sentence – although implausible – is actually false. For example, it is implausible but not impossible that in the example event *Frank has a broken leg. He calls the plumber*, Frank did (for unknown but conceivable reasons) call the plumber after breaking his leg. Plausibility comes into play only when there is uncertainty (e.g., Friedman & Halpern, 2001) and this uncertainty may prevent a clear rejection of implausible information. Thus, it may be more difficult to affirm implausible information (compared to plausible information) but not necessarily easier to reject it.[1] If this is the case, then one might find a different pattern for stimuli that describe events which are impossible rather than merely implausible (a terminology which Warren & McConnell, 2007, use to discriminate between violations of semantic vs. world knowledge), in which sentences describing impossible events evoke a clear negative response tendency which also leads to facilitation for negative responses. For this purpose, it would be useful to include an adequate neutral condition in future experiments to determine precisely the extent to which interference and facilitation contribute to the observed pattern.

It is important to note here that the present study was not aimed at contributing to the debate on whether there is a distinction between semantic and world or event knowledge, but rather focused on the specific question of whether event knowledge is used nonstrategically to assess plausibility during on-line comprehension. However, as outlined above, our paradigm offers a novel tool that might be useful to elucidate processing differences between different types of knowledge violations in future research.

5. Conclusion

In conclusion, our results suggest that plausibility monitoring is a routine, nonstrategic process that is invariably interwoven with language comprehension. As such, our findings are in line with Singer's (2006) proposal that the verification of linguistic messages is not dependent on an evaluative processing goal but "rather emerges from the fundamentals of the cognition of reading" (p. 589). In this way, our study elucidates an aspect of plausibility effects that has so far received relatively little attention, namely the extent to which these effects are nonstrategic and may reflect more than simple "processing costs" of implausible information: Rather, they point towards a highly purposeful monitoring process that promotes

[1] We would like to thank an anonymous reviewer for suggesting this possibility.

the accuracy and stability of the mental representations which are constructed during language comprehension (Schroeder et al., 2008).

Role of the funding source

The research reported in this article was supported by the German Research Association (Deutsche Forschungsgemeinschaft, DFG, grants RI 1100/4-1 and RI 1100/4-2).

Acknowledgements

We would like to thank Marc Grosjean, Julia Knoepke, and Sascha Schroeder for helpful discussions, David Nitz for technical assistance, and Claudia Beck, Ricarda Moll, Sarah Rom, Miriam Schwarzenthal, and Monika Wieland for their help in collecting data.

References

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59, 390-412.

Bates, D., Maechler, M., Bolker, B., 2011. Linear mixed-effects models using S4 classes [Software].R Foundation for Statistical Computing, Vienna, Austria.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Second ed. Erlbaum, Hillsdale, NJ.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 39, 175-191.

Friedman, N. & Halpern, J. Y., 2001. Plausibility measures and default reasoning. Journal of the ACM 48, 648-685.

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.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. Science 304, 438-441.

Hagoort, P., Van Berkum, J. J. A., 2007. Beyond the sentence given. Philosophical Transactions of the Royal Society. Series B: Biological Sciences 362, 801-811.

Jackendoff, R., 2002. Foundations of language: Brain, meaning, grammar, evolution. Oxford University Press, Oxford, England.

Just, M. A., Carpenter, P. A., 1980. A theory of reading: From eye fixations to comprehension. Psychological Review 87, 329-354.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., McRae, K., 2011. Event-based plausibility immediately influences on-line sentence comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition 37, 913-934.

McRae, K., Spivey-Knowlton, M. J., Tanenhaus, M. K., 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language 38, 283-312.

Morey, R. D., 2008. Confidence intervals from normalized data: A correction to Cousineau (2005). Tutorials in Quantitative Methods for Psychology 4, 61-64.

Nieuwland, M. S., Van Berkum, J .J. A., 2006. When peanuts fall in love: N400 evidence for the power of discourse. Journal of Cognitive Neuroscience 18, 1098-1111.

Rapp, D. N., 2008. How do readers handle incorrect information during reading? Memory & Cognition 36, 688-701.

Rayner, K., 1998. Eye-movements in reading and information processing: 20 years of research. Psychological Bulletin 124, 372-422.

Rayner, K., Warren, T., Juhasz, B. J., Liversedge, S. P., 2004. The effect of plausibility on eye movements in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition 30, 1290-1301.

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.

Schroeder, S., Richter, T., Hoever, I., 2008. Getting a picture that is both accurate and stable: Situation models and epistemic validation. Journal of Memory and Language 59, 237-259.

Singer, M., 2006. Verification of text ideas during reading. Journal of Memory and Language 54, 574-591.

Stanovich K. E., West R. F., 1981. The effect of sentence context on ongoing word recognition: Tests of a two-process theory. Journal of Experimental Psychology: Human Perception and Performance 7, 658–672.

Stanovich K. E., West R. F., 1983. On priming by a sentence context. Journal of Experimental Psychology: General 112, 1–36.

Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., Majewski, H., 2007. The Time Course of Plausibility Effects on Eye Movements in Reading: Evidence from Noun-Noun Compounds. Journal of Experimental Psychology: Learning, Memory, and Cognition 33, 1162-1169.

Stroop, J. R., 1935. Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18, 643-662.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition 31, 443-467.

Van Berkum, J. J. A., Hagoort, P., Brown, C. M., 1999. Semantic integration in sentences and discourse: Evidence from the N400. Journal of Cognitive Neuroscience 11, 657-671.

Van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., Brown, C. M., 2003. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. Cognitive Brain Research 17, 701-718.

Warren, T., McConnell, K., 2007. Investigating effects of selectional restriction violations and plausibility violation severity on eyemovements in reading. Psychonomic Bulletin & Review 14, 770-775.

Table 1

Norms for Plausibility (Mean Proportion of "Plausible" Judgments in %) and Predictability (Mean Cloze Value in %) of the selected Items

	Plausibility	Predictability
Condition	*M (SD)*	*M (SD)*
Plausible		
Predictable	97.85 (4.60)	75.22 (20.16)
Non-predictable	96.02 (5.66)	5.91 (9.40)
Implausible		
Predictable	4.10 (6.73)	1.45 (4.06)
Non-predictable	4.39 (6.00)	0.11 (0.89)

Table 2

Results (Means and Standard Deviations by Experimental Condition) of Experiment 1

Condition	Plausible		Implausible	
	RT	Error Rate	RT	Error Rate
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Predictable				
Positive Response	953 (280)	.029 (.057)	1034 (333)	.025 (.055)
Negative Response	873 (261)	.095 (.117)	927 (331)	.086 (.101)
Non-predictable				
Positive Response	1026 (293)	.048 (.086)	1171 (371)	.070 (.106)
Negative Response	995 (317)	.096 (.131)	1009 (310)	.113 (.140)

Note. Means and standard deviations are based on participants as units of observation.

Table 3

Results (Means and Standard Deviations by Experimental Condition) of Experiment 2

Condition	Plausible		Implausible	
	RT *M (SD)*	Error Rate *M (SD)*	RT *M (SD)*	Error Rate *M (SD)*
Predictable				
Positive Response	647 (143)	.019 (.054)	694 (198)	.033 (.067)
Negative Response	655 (151)	.024 (.059)	666 (152)	.032 (.083)
Non-predictable				
Positive Response	662 (149)	.059 (.095)	675 (179)	.029 (.067)
Negative Response	678 (173)	.036 (.072)	673 (153)	.021 (.058)

Note. Means and standard deviations are based on participants as units of observation.

Figure Captions

Figure 1. Mean correct response latency as a function of plausibility (plausible, implausible) and orthographical correctness (correct, incorrect) in the orthographical judgment task of Experiment 1. Error bars correspond to ±1 standard error of the mean computed for within-subjects designs (Morey, 2008).

Figure 2. Mean correct response latency as a function of plausibility (plausible, implausible) and required response (positive, negative) in the color judgment task of Experiment 2. Error bars correspond to ±1 standard error of the mean computed for within-subjects designs (Morey, 2008).

Figure 1

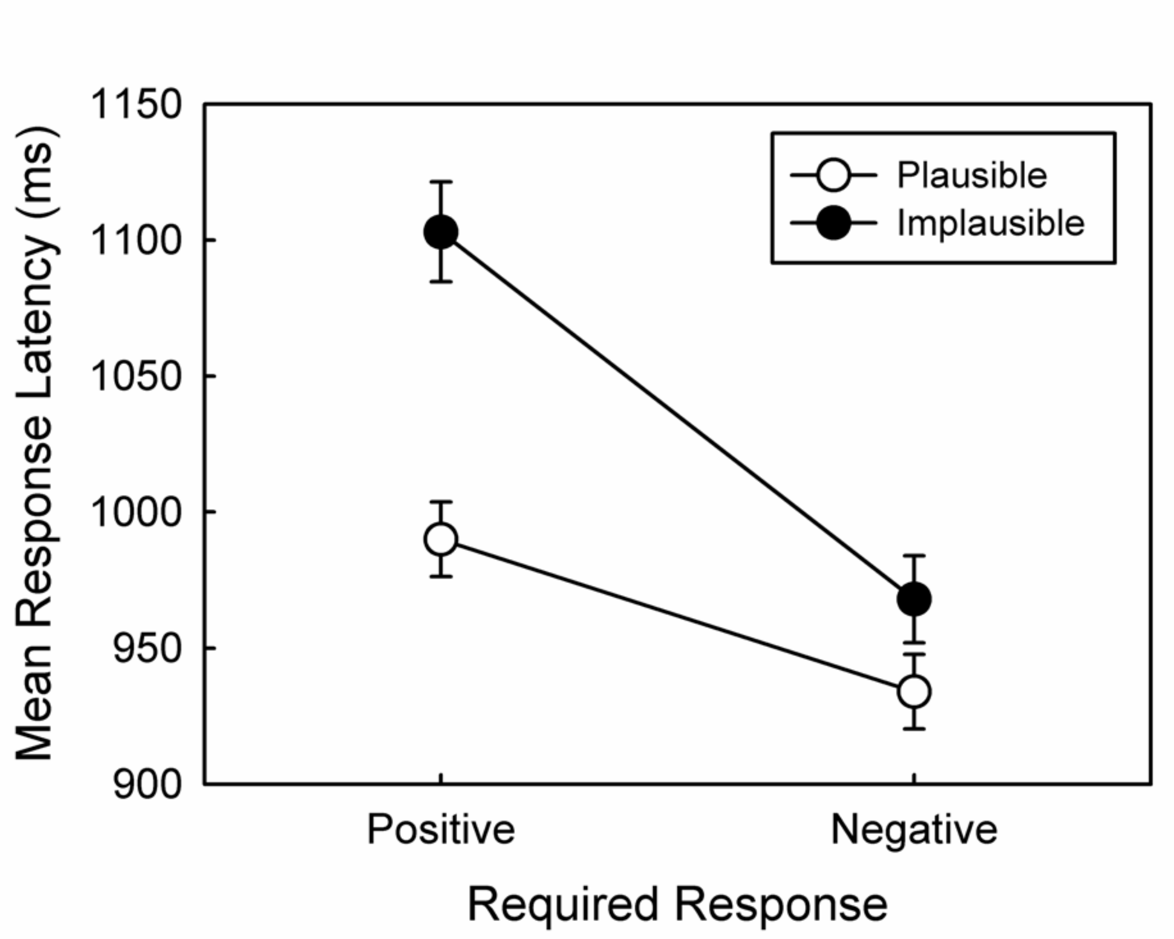
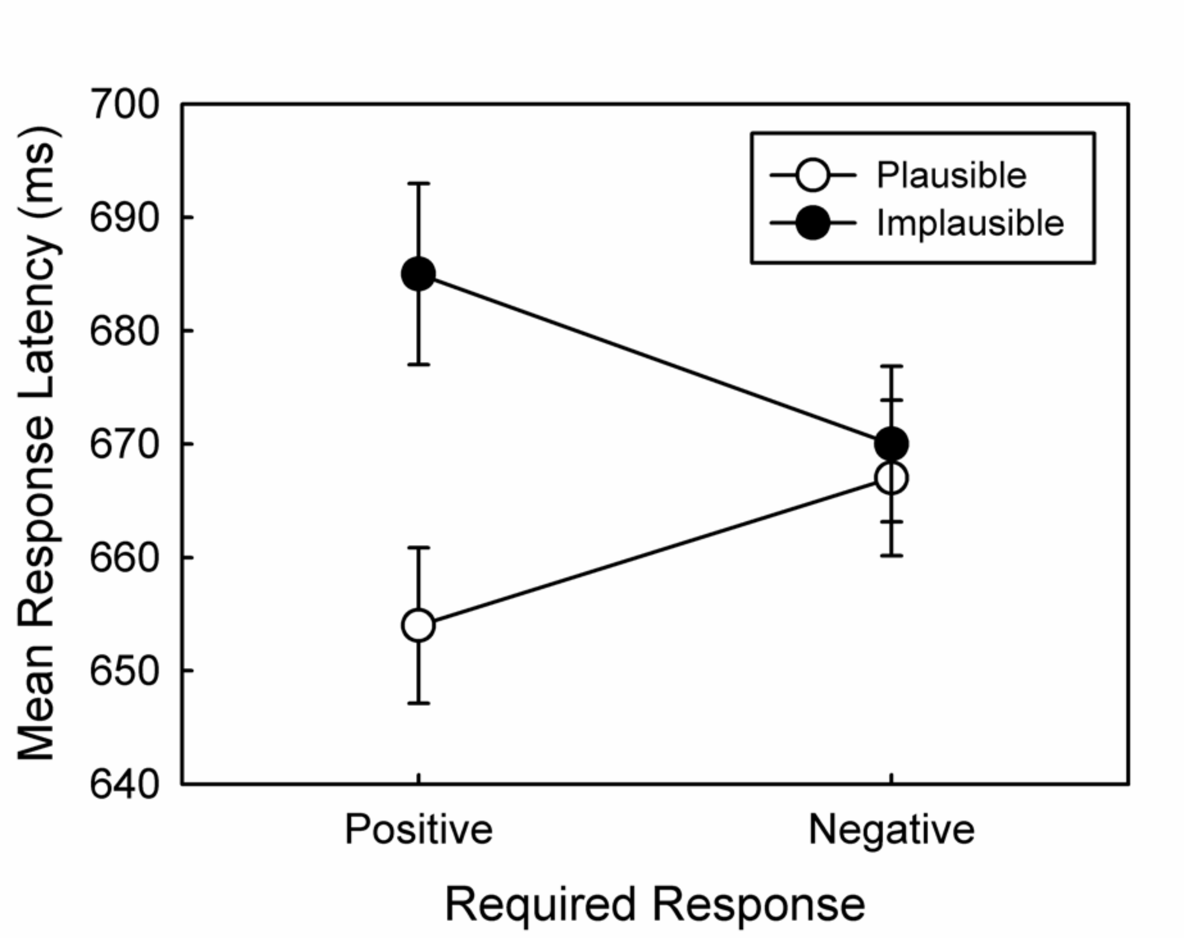

Mean Response Latency (ms)
1150
1100
1050
1000
950
900
Positive
Negative
Required Response
Plausible
Implausible

Figure 2


Mean Response Latency (ms)
700
690
680
670
660
650
640
Plausible
Implausible
Positive
Negative
Required Response