

Don't Believe Everything You Hear: Routine Validation of Audio-Visual Information in

Children and Adults

Accepted for publication in the journal *Memory & Cognition* (2018)

Benjamin A. Piest

Maj-Britt Isberner

University of Würzburg

University of Kassel

Tobias Richter

University of Würzburg

Author Note

This research was supported by the German Research Foundation (grant RI 1100/8-1, given to Tobias Richter). Experimental stimuli, data and R-scripts for the analyses are available from the third author upon request.

Corresponding author:

Tobias Richter

University of Würzburg, Department of Psychology IV

Röntgenring 10

97070 Würzburg, Germany

E-mail: tobias.richter@uni-wuerzburg.de

Telephone: +49-931 31 83755

Abstract

Previous research has shown that the validation of incoming information during language comprehension is a fast, efficient and routine process (epistemic monitoring). Previous research on this topic has focused on epistemic monitoring during reading. The present study extended this research by investigating epistemic monitoring of audio-visual information. In a Stroop-like paradigm, participants (Experiment 1: adults; Experiment 2: 10 year-old children) responded to the probe words “correct” and “false” by keypress after the presentation of auditory assertions that could be either true or false with respect to concurrently presented pictures. Results provide evidence for routine validation of audio-visual information. Moreover, the results show a stronger and more stable interference effect for children compared to adults.

Keywords: validation, epistemic Stroop effect, epistemic monitoring, audio-visual information, language comprehension

Many situations in which language is used not only require the comprehension but also the validation of incoming linguistic information – that is, judging whether the comprehended information is true or false. Recent studies investigating the relationship between comprehension and validation support the assumption that validation occurs immediately and routinely during language comprehension (e.g., Isberner & Richter, 2013, 2014a; Richter, Schroeder, & Wöhrmann, 2009; Singer, 2006). All of these studies have used written materials as stimuli, thus limiting the extant empirical evidence for routine validation to the domain of reading. However, if validation is an inherent component of language comprehension (Cook & O'Brien, 2014; O'Brien & Cook, 2016; Richter et al., 2009; Singer, 2013), it should not be restricted to the processing of written language. Spoken language is often used in face-to-face communications that are characterized by a richer pragmatic context, which includes the physical environment in which the communication is situated. This context potentially forms the basis for validation. Moreover, for communication to be successful, listeners need to align their mental representation with that of the speaker. The comprehension of definite expressions (e.g., sentences with demonstrative pronouns such as *This is a car*) is a case in point. For comprehending such expressions, listeners need to identify the intended referents that the speaker has in mind (Chafe, 1976). We propose that validation plays a crucial part in this process as it allows for monitoring the consistency of the content of a spoken message with the visual information that is in the focus of the listener's visual attention. In this way, validation of audio-visual information might play a major role in establishing and maintaining common ground (Clark & Brennan, 1992) during conversation.

In order to fulfil this function, validation of audio-visual information should proceed in a similarly passive and involuntary manner as the validation of written information. To test this assumption, we conducted two experiments using a Stroop-like paradigm adapted from

Isberner and Richter (2014a), one with adults and the other one with children. In the following, we will give a short overview of the theoretical background of our study and of previous research regarding language comprehension and validation.

Validation During Language Comprehension

Language comprehension involves more than the analysis of words, sentences, and texts. The meaning of a sentence must be integrated with information from prior sentences as well as with pertinent background knowledge. This integration process results in a mental representation of the state of affairs described in the text (situation model: van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). However, there is a growing consensus that comprehension involves routine and nonstrategic validation processes (O'Brien & Cook, 2016; Richter, 2015; Singer, 2013). Validation processes are assumed to check incoming information for inconsistencies and implausibility and have an effect on whether a particular piece of information becomes part of the current situation model or not (epistemic monitoring, Richter et al., 2009; Schroeder, Richter, & Hoever, 2008). Conscious and strategic validation of incoming information like assumed by two-step models of comprehension and validation (Gilbert, 1991) would be unsuited for this purpose. In contrast to the kind of validation highlighted in these models, epistemic monitoring processes are assumed to require little cognitive resources because they rely on knowledge that is activated through passive memory-based processes (e.g., McKoon & Ratcliff, 1995; Myers & O'Brien, 1998; O'Brien & Albrecht, 1992) and are themselves passive and involuntary. Therefore, no conscious and resource-demanding strategies are needed to activate the knowledge that incoming information is validated against, and no such strategies are needed to use the activated knowledge for validating incoming information.

A growing body of research supports the idea of routine and nonstrategic validation during language comprehension (e.g., Isberner & Richter, 2013, 2014a; O'Brien & Cook,

2016; Richter et al., 2009; Singer, 2006; see Isberner & Richter, 2014b, for an overview). Some of these studies have used an epistemic version of the Stroop paradigm (Stroop, 1935). In applications of this paradigm, participants judged whether the last word of a sentence was spelled correctly or incorrectly (Isberner & Richter, 2013; Richter et al., 2009), whether the last word of a sentence had changed color (Isberner & Richter, 2013), or which one of two possible probe words (correct or false) appeared immediately after the presentation of a sentence (Isberner & Richter, 2014a; a task originally introduced by Wiswede, Koranyi, Müller, Langner & Rothermund, 2013). The experimental sentences were either true (e.g., *Mountains are high*) or false (e.g., *Soft soap is edible*) and presented word-by-word at a fixed rate of presentation (e.g., 300ms/word). Importantly, the truth value of the sentences was irrelevant for responding to the focal task (i.e. orthographic judgments, color change judgments, or probe word identification), in analogy to the original Stroop task where the meaning of a word is irrelevant for the task of naming the color in which it is printed. One difference to the original Stroop task is the (sometimes) asynchronous presentation of the sentence that is validated and the stimulus for the focal task (e.g. the probe word "correct" or "false" that requires pressing one of two keys). However, it is important to note that the validation response can be formed only at the point where the truth-value of the sentence can be computed, and almost all applications of the paradigm (including the present experiments) have presented the stimulus for the focal task immediately after this point (the only exception being the experiment by Wiswede et al., 2013).

In all applications of the paradigm, a congruity effect between the validity of a sentence and the required response in the judgment or identification task occurred (epistemic Stroop effect: Richter et al., 2009). Participants showed slower response times for conditions in which the validity of the sentence and the required response in the task were incongruent compared to congruent conditions. For example, spelling judgments requiring the response

"yes" (Is the word spelled correctly?) were slower after invalid sentences (e.g., *Soft soap is edible*) compared to valid sentences (e.g. *Mountains are high*; Richter et al., 2009, Experiments 3 and 4). These results may be interpreted as supporting the idea of routine and nonstrategic validation processes during language comprehension, as participants were not able to ignore the validity of the sentences even when it was irrelevant to the task. Other studies using reading times or event-related potentials as indicators of validation (e.g., Ferretti, Singer, & Patterson, 2008; Singer, 2006) similarly support the idea of routine and nonstrategic validation of incoming information during language comprehension – although they, unlike studies using the epistemic Stroop paradigm, do not directly test the involuntary nature of validation (i.e., whether it can be suppressed if necessary). However, common to all of the abovementioned studies is that they are concerned with validation during reading, although the assumption of routine and nonstrategic validation processes, from a theoretical perspective, is not limited to comprehension in a specific modality. So far, there are no studies that have tested the assumption of routine validation processes during the comprehension of spoken language. Therefore, given the potential relevance of validation in face-to-face communications, one goal of the present study is to examine the epistemic Stroop effect (Richter et al., 2009) for oral language comprehension.

Integrating and Validating Linguistic Information with Visual Information

In many situations involving oral language comprehension, such as watching TV or engaging in conversations, the incoming auditory information is accompanied by information from the listener's visual environment. Very often comprehension requires processing these different sources of information in conjunction, which necessitates directing visual attention to (potential) referents of the linguistic input in the real world. Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) were among the first to investigate systematically the interplay between visual and linguistic information. They had participants listen to auditory sentences

like “Put the apple on the towel in the box”, in which one phrase (in this example: “on the towel”) was ambiguous (it could be either a modifier or a destination), while concurrently presenting them with sets of objects and recording their eye movements. The results showed that listeners presented with ambiguous sentences use visual information immediately to avoid a syntactic misanalysis (Tanenhaus et al., 1995). This paradigm used by Tanenhaus and colleagues is now known as the *visual world paradigm* and has inspired a large body of research during the last decades (for a detailed review, see Huettig, Rommers, & Meyer, 2011). In later experiments, the visual world paradigm has been used to show that people draw on both the visual context and their world knowledge to anticipate upcoming linguistic input (e.g., Altmann & Kamide, 1999, 2007, 2009; Knoeferle & Crocker, 2006, 2007). Altmann and Kamide (1999) presented participants with semi-realistic pictures of scenes (e.g., a boy, a cake, and toys) and auditory sentences regarding these scenes (e.g., “The boy will eat/move the cake”). They showed that participants started to focus significantly earlier on the cake in the “eat” condition compared to the “move” condition. Even though there was no explicit experimental instruction to focus on the screen or – as in the study by Tanenhaus et al. (1995) – to move a particular object, participants focused on the relevant objects mentioned in the sentence.

Studies using the visual world paradigm provide compelling evidence that comprehenders immediately and incrementally integrate linguistic and visual information, enabling them to both disambiguate and anticipate linguistic input. However, this also suggests the existence of an efficient mechanism to detect discrepancies between these two sources of information. We propose that validation serves as such a mechanism, just as it allows detecting discrepancies between linguistic information and prior knowledge. In the present study, we tested this assumption with a variant of the epistemic Stroop paradigm which combined pictures with auditorily presented sentences that either matched or

mismatched the content of the picture. The paradigm emulated a basic requirement in face-to-face communication, i.e. for listeners to monitor whether their visual attention is focused on the correct target of demonstrative referential expressions.

How Stroop-Like is the Epistemic Stroop Effect?

Most of the previous studies investigating validation during language comprehension using a Stroop-like paradigm have reported an interference effect between the required response in a simple judgment or decision task and the task-irrelevant truth value of an assertion presented immediately before (e.g., Isberner & Richter, 2013, 2014a; Richter et al., 2009). However, one may wonder whether this effect is indeed evidence for routine and nonstrategic validation processes or whether it could just be an artifact of the task that might induce an evaluative mindset (as criticized by Wiswede et al., 2013). To address this question, it is useful to examine how the epistemic Stroop effect compares to the classical Stroop effect (Stroop, 1935). The classical Stroop paradigm (Stroop, 1935) as a commonly used and well-evaluated paradigm has proven to be a good instrument to investigate routine and automatic processes. Although the Stroop effect is known to be a robust effect, studies have shown that participants can learn to control Stroop interference with practice (Dulaney & Rogers, 1994; Ellis & Dulaney, 1991; Ellis, Woodley-Zanthos, Dulaney, & Palmer, 1989). Therefore, the magnitude of the Stroop effect seems to be a function of the time on task. Thus, if the epistemic Stroop effect is indeed due to routine validation processes during language comprehension, the interference effect should be strongest at the beginning of the experiment and decrease over the course of the experiment. This prediction is rooted in the assumption that individuals will be able to learn to suppress the response tendency resulting from the validation process, as they are able to learn to suppress the reading response in the classical Stroop color-naming task.

Is the epistemic Stroop effect stronger for children than for adults?

Developmental studies investigating the classical Stroop effect reported an inverted U-shaped trajectory of Stroop interference. The Stroop effect appeared in elementary school children already able to read and increased during the first two to three years of reading practice, and then decreased continuously during adolescence (e.g., Comalli, Wapner, & Werner, 1962; Dash & Dash, 1982; Peru, Faccioli, & Tassinari, 2006; Rand, Wapner, Werner, & McFarland, 1963; Schadler & Thissen, 1981; Schiller, 1966). It has been suggested that the Stroop interference is a function of reading practice, but that this positive relationship is overshadowed by the concurrent development of inhibitory capacity, which improves during early adolescence (Comalli et al., 1962). That would explain the increase of Stroop interference in the first years after reading acquisition, and the decrease during adolescence. In line with this idea, a number of studies have shown larger Stroop interference for children than for adults (e.g., Carter, Mintun, & Cohen, 1995; Comalli et al., 1962; Guttentag & Haith, 1978; Vurpillot & Ball, 1979), and there are other studies suggesting that children have relatively weaker inhibitory control (Ridderinkhof, Band, & Logan, 1999; Tipper, Bourque, Anderson, & Brehaut, 1989; cp. Bub, Masson, & Lalonde, 2006). Therefore, if the validation of oral language against information from the visual context is indeed a routine process, similar age-related differences should occur for the audio-visual epistemic Stroop effect as well. Specifically, we expected children to exhibit a larger and more stable epistemic Stroop effect than adults. If the epistemic Stroop effect is already present (and even stronger) in children, this would also constitute evidence for the assumption that validation is not merely a learned higher-level reading process that becomes automatized over time, but indeed inherent to and a fundamental component of language comprehension.

Rationale of the Present Experiments

The present research aimed at answering three related research questions that revolve around the assumed routine validation of linguistic information. First, given that earlier research on validation focused on readers' knowledge-based validation in the processing of written information only, we sought to establish an epistemic Stroop effect with audio-visual information. We base this endeavor on the assumption that validation processes play an important role in aligning linguistic and visual information in communication situations. In two experiments, we tested this assumption by showing participants pictures of objects and concurrently presenting auditory assertions about these objects that were either true or false. After the presentation of each picture-assertion combination, participants responded to one of two probe words ("correct" or "false") by pressing one of two different keys. The probe words could be either congruent or incongruent with the truth-value of the picture-assertion combination (audio-visual epistemic Stroop-task). More specifically, the experimental sentences were simple assertions with a demonstrative that required a visual context for interpretation (e.g., *This is a car*). However, the task was not to interpret or validate the assertion or to relate its content to the picture, but only to identify the probe word that appeared after the assertion by pressing the corresponding key. For conditions in which the truth-value of the picture-assertion combination and the target word were incongruent, we expected slower responses to the target word and higher error rates compared to conditions in which the truth-value of the picture-assertion combination and the target word were congruent (audio-visual epistemic Stroop effect).

The second aim was to show that the audio-visual epistemic Stroop effect is strongest in the beginning of an experiment and decreases over the course of an experiment as the number of incongruent picture-assertion combinations that a participant has seen increases. This pattern would provide additional evidence for the notion of routine validation processes that occur spontaneously but whose interference can be suppressed through strategies learned

over the course of an experiment, just like the interference by the reading response in the original Stroop task.

The third aim was to investigate whether the magnitude of the audio-visual epistemic Stroop effect differs between children and adults. We conducted two experiments, Experiment 1 with adults and Experiment 2 with children (fourth-graders), to test the assumption that the epistemic Stroop effect is stronger for children compared to adults.

Experiment 1

Experiment 1 used the audio-visual version of the epistemic Stroop paradigm with adults. The trials were divided into three blocks to investigate potential changes in the magnitude of the epistemic Stroop effect over the course of the experiment.

Method

Participants. Sixty-nine undergraduate psychology students from the University of Kassel (Germany) participated in the experiment in exchange for course credit. All participants (54 females and 15 males) reported normal or corrected-to-normal vision and were either native German speakers or spoke German since the age of six. Their average age was 25.8 years ($SD = 7.1$).

Stimulus material. The stimuli were valid or invalid auditory assertions about simple pictures. All pictures depicted a colored or black object, for example a car, on white background. The pictures were simple, schematic pictures for which conflicts with participants' world knowledge were highly unlikely. The auditory assertions had the structure "This is [a/an] [concept noun]", for example *This is a car*, and were 2000 ms long with a flow time between 150 and 250 ms after the last audible sound. The final stimulus set consisted of 240 assertions and 240 pictures. Two pictures of the same category, for example *car/bike* of the category *vehicles*, and their corresponding valid assertions, for example *This is a car/This is a bike*, were combined to create an item with four versions (two valid and two invalid

versions, Figure 1). To make sure that no picture or assertion was presented more than once to the same participant, we created four lists with 120 items each (60 valid and 60 invalid versions) including one version of each item.

Norming study. The complete material was pilot-tested. The 14 participants of the norming study were asked to indicate for an original pool of 182 items whether the assertion about the picture was valid (correct) or invalid (incorrect). The items were presented in random order on a computer screen, using four different item lists to counterbalance the four versions of each item. All items with more than two incorrect responses across all four versions were dropped from the item pool. Furthermore, all items with response latencies that deviated more than three standard deviations from a participant's overall mean, a participant's mean reaction time for valid items or a participant's mean reaction time for invalid items were dropped. This resulted in a set of 121 items, 120 of which were selected as experimental items; 62 further items from the original item pool were used as example or icebreaker items.

Procedure. Participants were tested in groups of up to five people and instructed to press one of two keys in response to the probe words "correct" (German: richtig) or "false" (German: falsch). Participants responded to the probe word "correct" by pressing the key "K" with the index finger of the right hand and to the probe word "false" by pressing the key "D" with the index finger of the left hand. Thus, the validity of the picture-assertion combination was irrelevant for the probe word task. In 30 of the 120 trials, participants were prompted to categorize the object presented in the picture by pressing a key (four response options). These control questions were used to ensure that the participants paid attention to the pictures (and could not, for example, simply close their eyes to suppress the assumed interference). Here again, the validity of the picture-assertion combination was irrelevant to successfully solve the task.

The sequence of each trial was as follows: After a fixation point displayed for 500 ms, the picture was displayed for 2200 ms. One-hundred ms after the onset of the picture, the participants heard the auditory assertion about the picture via headphones. Each assertion had a length of 2000 ms. One-hundred ms after the offset of the assertion, the picture disappeared and one of the two probe words was presented until participants provided a response to the presented probe word (Figure 2). In the 30 trials that contained an additional categorization task, the prompt to categorize the picture and four alternative responses appeared on the screen immediately after the response to the probe word. The items were presented in three blocks of 40 items each. After each block, participants were allowed to take a short break before starting the next block. Furthermore, they received feedback regarding their response latencies and accuracy in the previous block. If a person's accuracy was lower than 80% in the previous block, the person was reminded that the task was not to validate the truth-value of the picture-assertion combination, but to simply respond correctly to the presented probe word. The experiment took on average between 20 and 30 minutes.

Design. The design was a 2 (validity: valid vs. invalid picture-assertion combination) x 2 (probe word: correct vs. false) x 3 (block: 1 vs. 2 vs. 3) within-subjects design. The dependent variables were the response latencies and the response accuracy in the epistemic Stroop task.

Results and Discussion

Response latencies and error rates were analysed with linear mixed effects models by using the `lmer` and `glmer` function of the R package `lme4` version 1.12 (Bates, Mächler, Bolker, & Walker, 2015). Interactions were further analysed using the `lsmeans` function in the `lsmeans` package (Lenth, 2016). In all significance tests, type-I-error probability was set at .05 (two-tailed).

Data cleaning. In a first step of data cleaning, responses within 10 ms after stimulus onset or exceeding 5 s were removed from the data set (0.07% of the data points). In the second step, participants and items were screened for unnaturally high error rates. Data from participants with an error rate of more than 40% in the epistemic Stroop task were removed from the data set, resulting in the exclusion of two participants. This cut-off was chosen because even though the task was an easy one, we expected participants to make more errors in incongruent conditions compared to congruent conditions. Thus, an error rate of 50% could be a result of random responses, a very strong epistemic Stroop effect, or a misunderstanding of the task. The cut-off was chosen to be low enough to exclude participants that responded randomly and high enough to keep participants that showed a strong epistemic Stroop effect. The average error rate for the experimental items in the epistemic Stroop task was 1.8%, with no item exceeding an error rate of 8%. Therefore, no items needed to be removed. All participants made less than 40% errors when responding to the control questions, and no participants were removed based on this criterion. Overall, the data cleaning resulted in a data set with 8038 data points. This data set was used for the analysis of the error rates.

For the analysis of the response latencies, a third step of data cleaning was applied to the data set of correct responses resulting from the first data cleaning. An inspection of the distribution revealed the positive skew typical for response latencies, which often leads to a non-normal distribution of the residuals (and thus, a violation of the assumptions of linear mixed models). To find the most adequate transformation for achieving a more symmetrical distribution, a Box-Cox analysis was used, revealing a lambda close to zero ($\lambda = -0.11$). Based on the ladder of powers (Mosteller & Tukey, 1977), a log-transformation was determined to be the most adequate transformation to reduce the non-normality. Response latencies deviating more than two standard deviations from the log-transformed mean of each

participant (4.5% of the data points) were treated as outliers and removed from the data set. This final step of data cleaning procedure resulted in a data set with 7545 data points.

Response latencies. Response latencies of correct responses were analyzed with a linear mixed model with random effects (random intercepts) of subjects and items.

Table 1 provides estimates and significance tests of the fixed effects. Here and in the remainder of the manuscript, we describe only the effects relevant for our hypotheses. (The other significant effects are displayed in the Tables 1 to 4; none of them affected the interpretation of hypothesis-relevant results.)

First of all, the predicted epistemic Stroop effect in terms of an interaction of probe word and validity of the picture-assertion combination was significant. Planned comparisons revealed that this interaction was due to responses to the probe word “false” ($M = 535$ ms, $SE = 10$ ms) being slower than responses to the probe word “correct” ($M = 515$ ms, $SE = 9$ ms) after valid picture-assertion combinations, $t(91.7) = -6.19, p < .001$, whereas no significant difference was observed for responses after invalid picture-assertions, $t(91.2) < 1$.

Furthermore, the three-way interaction of probe-word, validity, and block was significant (Figure 3). Separate follow-up tests for each block revealed a disordinal interaction between validity and probe word, $t(7443.8) = -4.43, p < .001$, in Block 1: After valid picture-assertion combinations, responses to the probe word “false” ($M = 553$ ms, $SE = 11$ ms) were slower compared to the probe word “correct” ($M = 536$ ms, $SE = 10$ ms), $t(91.4) = 2.83, p < .05$. For invalid picture-assertion combinations, the reverse effect occurred. Here, responses to the probe word “correct” ($M = 571$ ms, $SE = 11$ ms) were slower compared to the probe word “false” ($M = 550$ ms, $SE = 11$ ms), $t(92) = -3.31, p < .01$. Thus, responses were overall slower when the probe word mismatched than when it matched the task-irrelevant validity of the picture-assertion combination.

In Block 2, the follow-up analysis again revealed a significant two-way interaction of probe word and validity, $t(7443.8) = -3.41, p < .001$. This interaction was now semi-disordinal, with responses to the probe word “false” ($M = 531$ ms, $SE = 10$ ms) still being slower compared to responses to the probe word “correct” ($M = 509$ ms, $SE = 10$ ms) after valid picture-assertion combinations, $t(92.1) = 3.84, p < .01$. However, the difference between the probe words after invalid picture-assertion combinations disappeared, $t(91.7) = -.93, p = .787$.

In Block 3, the two-way interaction of probe word and validity was no longer significant, $t(7443.8) = -0.53$, n.s. Instead, there was a strong main effect of probe word, $t(7446.5) = 5.27, p < .001$, with responses to the probe word “false” ($M = 522$ ms, $SE = 10$) now being generally slower than to the probe word “correct” ($M = 499$ ms, $SE = 10$), regardless of the validity of the picture-assertion combination.

Error rates. The error rates in the epistemic Stroop task were analyzed with generalized linear mixed models with subjects and items included as random effects (random intercepts). Table 2 provides estimates and significance tests of the fixed effects.

First of all, the predicted epistemic Stroop effect, i.e. the interaction of probe word and validity of the picture-assertion combination, only showed a tendency in the predicted direction and did not become significant, $z = -1.77, p < .1$. However, planned comparisons revealed that as predicted, the probability of false responses to the probe word “correct” (probability = .02, $SE = .00$) was slightly higher than for the probe word “false” (probability = .01, $SE = .00$) after invalid picture-assertion combinations, $z = -3.01, p < .01$. After valid picture-assertion combinations, error probability did not differ between the probe words “false” (probability = .01, $SE = .00$) and “correct” (probability = .01, $SE = .00$), $z = -0.51$, n.s. The three-way interaction of probe word and validity with block was not significant. Separate follow-up tests for each block yielded no significant interaction of probe word and validity.

Only in Block 1, the interaction between probe word and validity approached significance, $z = -1.92$, $p < .1$, again due to a higher error probability for the probe word “correct” (probability = .02, $SE = .01$) compared to the probe word “false” (probability = .00, $SE = .00$) after invalid picture-assertion combinations, $z = -2.64$, $p < .01$ (see also Figure 4).

In sum, these results support the assumption of routine validation of audio-visual information. An epistemic Stroop effect occurred for the response latencies in the overall analyses across all three blocks of the experiment (and a similar, though not significant, pattern emerged for the error rates). Furthermore, separate analyses for each block showed that the epistemic Stroop effect was present from the very beginning of the experiment. However, it looks like participants were able to develop strategies against the interference of the task-irrelevant validity of the picture-assertion combinations with the task of responding to the probe words. At the beginning of the experiment, participants showed a symmetrical epistemic Stroop effect in the response latencies with slower responses to validity-incongruent probe words after both valid and invalid picture-assertion combinations. However, this effect decreased in Block 2 and was no longer significant in Block 3.

These results suggest that adults are able to avoid the interference by strategically inhibiting the response tendency resulting from the validation process, which somewhat distorts the overall effect. For this reason, in Experiment 2, we applied the paradigm used in Experiment 1 to children in Grade 4. Due to the weaker inhibitory capacity in this population (e.g., Carter et al., 1995; Comalli et al., 1962; Guttentag & Haith, 1978; Ridderinkhof et al., 1999; Tipper et al., 1989; Vurpillot & Ball, 1979) compared to the adult participants of Experiment 1, we expected the audio-visual epistemic Stroop effect to be even stronger and more stable in Experiment 2, while the general pattern of results should be preserved.

Experiment 2

Method

Participants. Two hundred and eighty-three fourth-graders of different primary schools in the Kassel (Germany) region participated in the experiment. All participants (123 girls and 154 boys, no gender data provided for 6 children) were either native German speakers or spoke German since the age of six. Their average age was 10.5 years ($SD = 0.5$, no age data provided for 49 children).

Stimulus material. The pictures and the assertions were the same as in Experiment 1. The only difference was that the instructions were adapted to be suitable for children and that the example trial feedback on why a response was correct or incorrect was more detailed.

Procedure. The procedure was similar to Experiment 1, except for the fact that participants were tested in groups of up to 24 people in school classrooms. Moreover, the assignment of the response keys to the probe words was slightly different: Participants responded to the probe word “correct” by pressing the key “J” and to the probe word “false” by pressing the key “F”.

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

Results and Discussion

Data cleaning. The data cleaning procedure was the same as in Experiment 1. Responses within 10 ms of stimulus onset or exceeding 5 s were removed from the data set (2.1% of the data points). Next, participants with an error rate of more than 40% in the control questions were removed from the data set, which resulted in the exclusion of 30 participants. Furthermore, all participants with an error rate of more than 40% in the epistemic Stroop task were removed from the data set, which resulted in the exclusion of another 61 participants. The average error rate for the experimental items was 10.1%, with no item exceeding an error rate of 16.4%. Therefore, no items needed to be removed. This

general data cleaning procedure resulted in a data set with 20,312 data points. This data set was used for the analysis of the error rates.

For the data cleaning of the response latencies of the correct responses, the Box-Cox analysis revealed again a lambda close to zero ($\lambda = 0.15$). Therefore, the response latencies were log-transformed and response latencies deviating more than two standard deviations from the log-transformed mean of each participant (5.3% of the data points) were treated as outliers and removed from the data set. This data cleaning procedure resulted in a data set with 17,288 data points.

Response latencies. Like in Experiment 1, we estimated a linear mixed effects model for the response latencies of the correct responses in the Stroop task with random effects (random intercepts) of subjects and items (Table 3). Importantly, there was again a significant two-way interaction of probe word and validity of the picture-assertion combination. Planned comparisons showed that after invalid picture-assertion combinations, participants were slower to respond to the probe word “correct” ($M = 808$ ms, $SE = 16$ ms) than to the probe word “false” ($M = 779$ ms, $SE = 16$ ms), $t(301.6) = 4.1, p < .001$. For valid picture-assertion combinations, the effect was reversed and even stronger. Here, participants were slower to respond to the probe word “false” ($M = 818$ ms, $SE = 16$ ms) than to the probe word “correct” ($M = 733$ ms, $SE = 15$ ms), $t(305.6) = 11.8, p < .001$.

Furthermore, the three-way interaction of all three independent variables was significant (Figure 5). Separate follow-up analyses for each block revealed a disordinal interaction between validity and probe word for Block 1, $t(17074.7) = -10.28, p < .001$. After valid picture-assertion combinations, responses to the probe word “false” ($M = 853$ ms, $SE = 19$ ms) were slower than responses to the probe word “correct” ($M = 733$ ms, $SE = 16$ ms), $t(305.6) = 9.7, p < .001$, whereas after invalid picture-assertion combinations, responses to the probe word “correct” ($M = 852$ ms, $SE = 19$ ms) were slower than responses to the probe

word “false” ($M = 777$ ms, $SE = 17$ ms), $t(301.6) = -6.0$, $p < .001$. In Block 2, this disordinal interaction between validity and probe word decreased but remained significant, $t(17074.7) = -5.47$, $p < .001$. After valid picture-assertion combinations, responses to the probe word “false” ($M = 798$ ms, $SE = 17$ ms) were slower compared to the probe word “correct” ($M = 732$ ms, $SE = 16$ ms), $t(316) = 5.7$, $p < .001$. After invalid picture-assertion combinations, responses to the probe word “correct” ($M = 796$, $SE = 17$ ms) were slower compared to the probe word “false” ($M = 763$ ms, $SE = 17$ ms), $t(314.4) = -2.9$, $p < .05$. In Block 3, the interaction between validity and probe word was again reduced, $t(17074.7) = -2.23$, $p < .001$, and now only semi-disordinal. The difference between responses to the probe words after invalid picture-assertion combinations was no longer significant, $t(355.2) = 1.5$, $p = .417$, whereas the effect for valid picture-assertion combinations with slower responses to the probe word “false” ($M = 803$ ms, $SE = 18$ ms) than to the probe word “correct” ($M = 734$ ms, $SE = 16$ ms) remained stable, $t(351.9) = 5.4$, $p < .001$.

Error rates. Like in Experiment 1, we estimated a generalized linear mixed effects model for the error rates with random effects (random intercepts) of subjects and items (Table 4). Most importantly, we found an interaction between probe word and validity of the picture-assertion combination. This epistemic Stroop effect was driven by the fact that after invalid picture-assertion combinations, participants had a higher probability of responding erroneously to the probe word “correct” (probability = .08, $SE = .01$) than to the probe word “false” (probability = .03, $SE = .00$), $z = 14.2$, $p < .001$, whereas for valid picture-assertion combinations, the effect was reversed: Here, participants were more likely to make errors in responding to the probe word “false” (probability = .08, $SE = .01$) than in responding to the probe word “correct” (probability = .02, $SE = .00$), $z = 16.9$, $p < .001$.

Furthermore, and paralleling the analysis for the response latencies, the analysis revealed a significant three-way interaction between validity, probe word, and block, driven

by the pattern that the epistemic Stroop effect continually decreased over the three blocks (Figure 6). We followed up on this three-way interaction with separate analyses for each block. For Block 1, the analysis revealed a strong disordinal interaction between probe word and validity, $z = -21.8, p < .001$. After the presentation of a valid picture-assertion combination, error probability was much higher for the probe word “false” (probability = .14, $SE = .02$) than for the probe word “correct” (probability = .02, $SE = .00$), $z = 15.5, p < .001$. Conversely, after the presentation of an invalid picture-assertion combination, error probability was much higher for the probe word “correct” (probability = .15, $SE = .02$) than for the probe word “false” (probability = .02, $SE = .00$), $z = 15.5, p < .001$. In Block 2, the same interaction emerged, but it was less strong than in Block 1, $z = -13.5, p < .001$. After valid picture-assertion combinations, error probability for the probe word “false” (probability = .09, $SE = .01$) was again higher than for the probe word “correct” (probability = .03, $SE = .00$), $z = 9.8, p < .001$, whereas for invalid picture-assertion combinations, error probability for the probe word “correct” (probability = .09, $SE = .01$) was again higher than for the probe word “false” ($M = .03, SE = .00$), $z = -9.4, p < .001$. Thus, in Block 2, the epistemic Stroop effect was reduced but still clearly present. In Block 3, the interaction decreased again in magnitude but remained significant, $z = -4.92, p < .001$. After valid picture-assertion combinations, error probability for the probe word “false” (probability = .05, $SE = .01$) was still significantly higher than for the probe word “correct” (probability = .02, $SE = .00$), $z = 5.1, p < .001$. After invalid picture-assertion combinations, error probability for the probe word “correct” (probability = .04, $SE = .01$) also remained higher than for the probe word “false” (probability = .03, $SE = .00$) but the difference just failed to reach significance, $z = -1.9, p < .1$. In sum, the epistemic Stroop effect decreased but the overall pattern was stable across blocks.

The results basically replicate the findings of Experiment 1 and also support the assumption of a stronger and more stable epistemic Stroop effect in children compared to adults. In contrast to the results for the adults, the epistemic Stroop effect remained significant across all three blocks in both the error rates and the response latencies, despite becoming weaker over time. Thus, children too seem to learn to inhibit the interference of the task-irrelevant truth of the picture-assertion combinations with responses in the probe task, albeit less successfully so than adults.

Joint analysis of Experiments 1 and 2

We conducted a joint linear mixed model analysis for the response time data of Experiments 1 and 2 and a joint generalized linear mixed model analysis for the error rate data of both Experiments, with validity (contrast coded: invalid = 1, valid = -1), probe word (contrast coded: false = 1, correct = -1) and experiment (contrast coded: adults = 1, children = -1) as fixed effect factors for each analysis.

Response latencies. As in the separate analysis of the response latencies in the two experiments, the two-way interaction between probe word and validity of the picture-assertion combination was significant, $t(24565.5) = -9.35, p < .001$. After a valid picture-assertion combination, the responses to the probe word “false” ($M = 677$ ms, $SE = 11$) were slower compared to the probe word “correct” ($M = 624$ ms, $SE = 10$), and after an invalid picture-assertion combination, the responses to the probe word “correct” ($M = 672$ ms, $SE = 11$) were slower compared to the probe word “false” ($M = 657$ ms, $SE = 11$). Thus, an epistemic Stroop effect occurred in the joint analysis of the two experiments. Furthermore, the three-way interaction between probe word, validity of the picture-assertion combination and Experiment was significant, $t(24497.9) = 5.41, p < .001$, indicating that the audio-visual epistemic Stroop effect was stronger in children compared to adults.

Error rates. The joint analysis of the error rates in Experiments 1 and 2 revealed a parallel pattern of results. First, an epistemic Stroop effect emerged, as indicated by a significant disordinal two-way interaction of probe word and validity of the picture-assertion combination, $z = -9.59, p < .001$. After a valid picture-assertion combination the error probability was higher for the probe word “false” (probability = .05, $SE = .01$) compared to the probe word “correct” (probability = .02, $SE = .00$). After an invalid picture-assertion combination the reverse effect occurred with a higher error probability for the probe word “correct” (probability = .06, $SE = .01$) compared to the probe word “false” (probability = .02, $SE = .00$). Furthermore, as in the analysis of the response latencies, the three-way interaction of probe word, picture-assertion combination and Experiment was significant, $z = 6.35, p < .001$, indicating that the audio-visual epistemic Stroop effect for the error rates was stronger in children compared to adults.

General Discussion

In this study, we assumed that individuals would show the same Stroop-like interference effects for audio-visual stimuli (spoken assertions matching or mismatching the content of pictures) that had been found by Richter et al. (2009) and Isberner and Richter (2014a) for visually presented assertions (matching or mismatching common world knowledge). Furthermore, we assumed that this interference effect would be present from the very beginning of the experiment and gradually decrease over its course, and that children (fourth-graders) would show a stronger effect compared to adults (university students). In order to test these assumptions, we used a Stroop-like paradigm adapted from Isberner and Richter (2014a). Participants responded to the probe words “correct” or “false” by key presses immediately after the presentation of valid or invalid auditory assertions about concurrently presented pictures.

In line with our assumption of a routine, passive validation of audio-visual information, the results showed an overall Stroop-like interaction effect of validity and probe word in both experiments, with longer response latencies and higher error rates when the probe mismatched the task-irrelevant validity. Thus, for the first time, evidence could be obtained that the passive and involuntary validation of linguistic information is not restricted to written information.

In addition, the follow-up analyses this interaction in each block shed light on the development of the epistemic Stroop effect over the course of the experiment in both groups. In line with the idea that the epistemic Stroop effect captures an automatic validation process and is not induced by properties of the task, the interaction effect was present from the very beginning of the experiment in both groups. In fact, in all analyses, the effect was most pronounced in the first block, but became weaker over the course of the experiment, which suggests that both adults and children learned to inhibit the response tendency resulting from the validation process. The answers of adults (Experiment 1) to the question of whether they used any strategy in the task support this assumption. About 25% of the participants of Experiment 1 responded something like: “I focused only on the words and tried to ignore the pictures and the assertions”. However, the stronger and more stable effects for children indicate that adults were both faster and more successful in inhibiting the response tendency resulting from the validation process. Moreover, the block-analysis shows that – except for the error rates in Experiment 1 (adults) – the pattern of the interaction is initially quite symmetrical for valid and invalid assertions and then tends to become asymmetrical as the interaction is overshadowed by increasingly stronger main effects. This results in an overall asymmetrical pattern of the interaction across blocks, similar to patterns found in previous experiments on the epistemic Stroop effect (Isberner & Richter, 2013, 2014a; Richter et al., 2009). The results of the current experiments thus indicate that such patterns most likely

result from two opposing processes: The nonstrategic validation process, which is initially dominant, and a (probably) strategic inhibition process, which becomes stronger over time.

It is important to note that even if participants learned to inhibit the response tendency resulting from the validation process, this does in no way contradict the assumption that validation is an automatic process. In fact, the decrease of the magnitude of the audio-visual epistemic Stroop effect over the course of the experiment mirrors findings on the classical Stroop effect (Dulaney & Rogers, 1994; Ellis & Dulaney, 1991; Ellis et al., 1989). In these studies, participants learn to inhibit the response tendency resulting from reading the word, which reduces the interference effect on color naming in incongruent trials. However, the training leaves the automaticity of reading processes intact. In much the same way, we would maintain that validation is an automatic process even if the magnitude of the epistemic Stroop effect can be reduced through training.

The reduction of the Stroop effect might also be due to participants having routinized their responses to the probe task during the experiment, making the responses more efficient and thus less susceptible to interference due to an incongruent response tendency resulting from the validation process. This processing speed account has originally been proposed as an explanation for training effects with the classical Stroop task (McLeod & Dunbar, 1988), and it is likely to hold for the epistemic Stroop task, too. In line with the processing speed account, participants responded generally faster in the probe task the longer the experiment lasted, which indicates routinization of this task. While the processing speed account might well explain parts of the results, Experiment 2 with children in fourth-grade and even more so, the comparison of these results with those of Experiment 1 with adults, suggests that the inhibition of the response tendency resulting from the validation process plays a role, too. As expected, children showed a larger overall epistemic Stroop effect in both dependent variables, but particularly in the error rates. This pattern of effects is probably due to the

weaker inhibitory capacity of children compared to adults (e.g., Carter et al., 1995; Comalli et al., 1962; Diamond, 2013; Guttentag & Haith, 1978; Ridderinkhof et al., 1999; Tipper et al., 1989; Vurpillot & Ball, 1979): It seems that the fourth-graders in Experiment 2 easily comprehended the spoken assertions but had more difficulties than adults to suppress the response tendency resulting from the validation process that accompanies comprehension. Inhibitory control, i.e., the ability to inhibit irrelevant information, evolves during childhood and adolescence (Bedard et al., 2002; Klenberg, Korkman, & Lahti-Nuutila, 2001). It would be interesting to investigate in future studies whether the magnitude of the epistemic Stroop interference decreases to the extent that inhibitory control increases from childhood to young adulthood. In any case, the development of inhibitory control might explain why in the present studies, an epistemic Stroop effect in the error rates did not reach significance in Experiment 1 (adults) but was present in all three blocks in Experiment 2 (children).

It must be noted that in terms of absolute mean differences, the audio-visual epistemic Stroop (based on the congruency of auditory language input with the current visual context) seems to be smaller than the epistemic Stroop effect based on the congruency of written language input with general world knowledge that Isberner and Richter (2014a) found using the same task as in the present experiments with an adult sample.

Interestingly, the results also suggest differences in the patterns for positive and negative responses and their changes across trials. Although these differences were not part of our hypotheses, a post-hoc comparison with previous studies suggests that they may indeed not be spurious. The overall pattern that emerges when comparing the results with those of previous epistemic Stroop experiments (Isberner & Richter, 2013, 2014a; Richter et al., 2009) is a stronger and/or more stable effect for positive than for negative responses in the response latencies, but a stronger and/or more stable effect for negative than for positive responses in the error rates. It thus seems that positive and negative validation responses

manifest differently, which could indicate differences in either their strength or their time course. Future studies may more specifically investigate this issue, which could provide important insights into the mechanisms underlying validation.

Previous research has focused on validation of information during reading and on how routine and fast these validation processes are (e.g., Isberner & Richter, 2013, 2014a; Richter et al., 2009). The purpose of the present research was to show that individuals validate audio-visual information in the same routine and fast manner (Richter et al., 2009). The results provide strong support for the assumption that situated verbal language comprehension entails a routine, nonstrategic, and efficient validation process which monitors the integration of verbal (assertion) and visual (picture) information within a specific situation. The RI-Val Model put forward by Cook and O'Brien (2014) has integrated the idea of validation as a process monitoring the integration of different sources of information. They propose that resonance (activation), integration, and validation are parallel but asynchronous processes that once started run to completion, with each stage activating the next stage. The integration process links incoming information to the contents of working memory. The validation process checks the result of the integration process against earlier parts of the text, a person's world knowledge, and the relation of both (Cook & O'Brien, 2014). Going beyond previous studies, the present study has shown that this validation process also monitors the fit of incoming verbal information with its (visual) context, or more precisely, the contextual truth of linguistic input. This monitoring process might serve an important function for face-to-face communications in helping communicators to establish and maintain a common ground (Clark & Brennan, 1991). More precisely, when speakers talk about aspects of the physical environment, listeners need to determine the correct referents (Chafé, 1976). As part of this process, they need to monitor constantly whether they focus their visual attention on the aspects of the situation that the speaker has in mind. In this case, information is activated via

perception (rather than resonance) and in the integration stage, links are formed between objects in the visual context and referents in the auditory language input. While the precise mechanism underlying the validation process is beyond the scope of this study, we would assume that the failure to establish (meaningful) links in this stage is tantamount to a negative validation response and that integration and validation are, in this sense, “two sides of the same coin” rather than separable stages (for an elaboration of this argument, see Richter, 2015).

The epistemic Stroop paradigm is, to our knowledge, so far the only paradigm that directly allows testing whether comprehension can be performed without validation when this is required by the task (and the results suggest that it cannot). Methodologically, however, the findings of the present study that the epistemic Stroop effect decreased during the experiments also suggest a limitation of the paradigm for the study of passive validation processes. Researchers who wish to use this paradigm should keep in mind that effects obtained with this paradigm might be reduced due to (probably) strategic inhibition, and that such inhibition might increase over the course of the experiment. One potential measure against these biases is to reduce the length of experiments because apparently, participants need a large number of trials to learn to inhibit the response tendency resulting from the validation process. Another reasonable measure would be to increase the number or variations of response options used during the experiment (e.g. using not only the probe "true" and "false", but also "plausible", "implausible", "correct" or "wrong").

In conclusion, the present study extends previous research on routine validation during language comprehension to audio-visual information and provides insight into possible moderating factors of the epistemic Stroop effect. Furthermore, it strongly supports the assumption that validation is not restricted to reading, but an obligatory component of language comprehension in general, serving spoken language comprehension in face-to-face

communications by checking the consistency of the linguistic message with the visual context. Nevertheless, further research is necessary to gain a better understanding of the underlying cognitive processes of validation during language comprehension.

References

- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264. doi:10.1016/S0010-0277(99)00059-1
- Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*, 502–518. doi:10.1016/j.jml.2006.12.004
- Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*, 55–71. doi:10.1016/j.cognition.2008.12.005
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01
- Bedard, A.-C., Nichols, S., Barbosa, J. A., Schachar, R., Logan, G. D., & Tannock, R. (2002). The development of selective inhibitory control across the life span. *Developmental Neuropsychology*, *21*, 93–111. doi:10.1207/S15326942DN2101_5
- Bub, D. N., Masson, M. E. J., & Lalonde, C. E. (2006). Cognitive control in children: Stroop interference and suppression of word reading. *Psychological Science*, *17*, 351–357. doi:10.1111/j.1467-9280.2006.01710.x
- Carter, C. S., Mintun, M., & Cohen, J. D. (1995). Interference and facilitation effects during selective attention: An H2150 PET study of Stroop task performance. *NeuroImage*, *2*, 264–272.
- Chafé, W.L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C.N. Li (Ed.), *Subject and topic* (pp. 26-55). New York: Academic Press.

- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L.B. Resnick, J.M. Levine, & J.S.D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.
- Comalli, P. E., Wapner, S., & Werner, H. (1962). Interference effects of Stroop color-word test in childhood, adulthood, and aging. *The Journal of Genetic Psychology, 100*, 47–51.
doi:10.1080/00221325.1962.10533572
- Cook, A. E., & O'Brien, E. J. (2014). Knowledge activation, integration, and validation during narrative text comprehension. *Discourse Processes, 51*, 26–49.
doi:10.1080/0163853X.2013.855107
- Dash, J., & Dash, S. (1982). Cognitive developmental studies of the Stroop phenomena: Cross-sectional and longitudinal data. *Indian Psychologist, 1*, 24–33.
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168.
doi:10.1146/annurev-psych-113011-143750
- Dulaney, C. L., & Rogers, W.A. (1994). Mechanisms underlying reduction in Stroop interference with practice for young and old adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 470-484. doi:10.1037/0278-7393.20.2.470
- Ellis, N. R., & Dulaney, C. L. (1991). Further evidence for cognitive inertia of persons with mental retardation. *American Journal on Mental Retardation, 95*, 613-621.
- Ellis, N. R., Woodley-Zanthos, P., Dulaney, C. L., & Palmer, R. L. (1989). Automatic-effortful processing and cognitive inertia in persons with mental retardation. *American Journal on Mental Retardation, 93*, 412-423.
- Ferretti, T. R., Singer, M., & Patterson, C. (2008). Electrophysiological evidence for the time-course of verifying text ideas. *Cognition, 108*, 881-888. doi:10.1016/j.cognition.2008.06.002
- Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*, 577–609.

- Guttentag, R. E., & Haith, M. M. (1978). Automatic processing as a function of age and reading ability. *Child Development, 49*, 707-716.
- Huetting, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*, 151–171. doi:10.1016/j.actpsy.2010.11.003
- Isberner, M.-B., & Richter, T. (2013). Can readers ignore implausibility? Evidence for nonstrategic monitoring of event-based plausibility in language comprehension. *Acta Psychologica, 142*, 15–22. doi:10.1016/j.actpsy.2012.10.003
- Isberner, M.-B., & Richter, T. (2014a). Does validation during language comprehension depend on an evaluative mindset? *Discourse Processes, 51*, 7–25. doi:10.1080/0163853X.2013.855867
- Isberner, M.-B. & Richter, T. (2014b). Comprehension and validation: Separable stages of information processing? A case for epistemic monitoring in language comprehension. In D.N. Rapp & J. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 245–276). Boston, MA: MIT Press.
- Klenberg, L., Korkman, M., & Lahti-Nuuttila, P. (2001). Differential development of attention and executive functions in 3- to 12-year-old Finnish children. *Developmental Neuropsychology, 20*, 407–428. doi:10.1207/S15326942DN2001_6
- Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science, 30*, 481–529. doi:10.1207/s15516709cog0000_65
- Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language, 57*, 519–543. doi:10.1016/j.jml.2007.01.003

- Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1). doi:10.18637/jss.v069.i01
- McKoon, G., & Ratcliff, R. (1995). The minimalist hypothesis: Directions for research. In C. A. Weaver, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 97–116). Hillsdale, NJ, Erlbaum.
- MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 126-135.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Pearson.
- Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*, 131–157. doi:10.1080/01638539809545042
- O'Brien, E. J., & Albrecht, J. E. (1992). Comprehension strategies in the development of a mental model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 777–784. doi:10.1037/0278-7393.18.4.777
- O'Brien, E.J., & Cook, A.E. (2016). Coherence threshold and the continuity of processing: The RI-Val Model of comprehension. *Discourse Processes*, *53*, 326–338. doi:10.1080/0163853X.2015.1123341
- Peru, A., Faccioli, C., & Tassinari, G. (2006). Stroop effects from 3 to 10 years: The critical role of reading acquisition. *Archives Italiennes de Biologie*, *144*, 45–62. doi:10.4449/aib.v144i1.896
- Rand, G., Wapner, S., Werner, H., & McFarland, J. H. (1963). Age differences in performance on the Stroop Color-Word test. *Journal of Personality*, *31*, 534–558. doi:10.1111/j.1467-6494.1963.tb01318.x

- Richter, T. (2015). Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes*, *52*, 337–355. doi:10.1080/0163853X.2015.1025665
- 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. doi:10.1037/a0014038
- Ridderinkhof, K. R., Band, G. P. H., & Logan, G. D. (1999). A study of adaptive behavior: Effects of age and irrelevant information on the ability to inhibit one's actions. *Acta Psychologica*, *101*, 315–337.
- Schadler, M., & Thissen, D. M. (1981). The development of automatic word recognition and reading skill. *Memory and Cognition*, *9*, 132–141. doi:10.3758/BF03202327
- Schiller, P. H. (1966). Developmental study of color-word interference. *Journal of Experimental Psychology*, *72*, 105–108. doi:10.1037/h0023358
- 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*, 237–255. doi:10.1016/j.jml.2008.05.001
- Singer, M. (2006). Verification of text ideas during reading. *Journal of Memory and Language*, *54*, 574–591. doi:10.1016/j.jml.2005.11.003
- Singer, M. (2013). Validation in reading comprehension. *Current Directions in Psychological Science*, *22*, 361–366. doi:10.1177/0963721413495236
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. doi:10.1037/h0054651
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634. doi:10.1126/science.7777863

- Tipper, S. P., Bourque, T. A., Anderson, S. H., & Brehaut, J. C. (1989). Mechanisms of attention: A developmental study. *Journal of Experimental Child Psychology*, *48*, 353-378.
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
- Vurpillot, E., & Ball, W. A. (1979). The concept of identity and children's selective attention. In G. A. Hale & M. Lewis (Eds.), *Attention and cognitive development* (pp. 23-42). New York: Plenum.
- Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (2013). Validating the truth of propositions: Behavioral and ERP indicators of truth evaluation processes. *Social Cognitive and Affective Neuroscience*, *8*, 647–653. doi:10.1093/scan/nss042
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185. doi:10.1037/0033-2909.123.2.162

Table 1:

Estimated Coefficients, Standard Errors, Degrees of Freedom, and t-values for the Linear Mixed Model of the Log-transformed Response Latencies in Experiment 1.

	<i>Est.</i>	<i>SE</i>	<i>df</i>	<i>t</i>	
(Intercept)	6.273	0.018	66.7	356.35	***
Probe word	0.009	0.002	7383.1	4.00	***
Validity	0.011	0.002	7378.3	4.79	***
Block 1	0.041	0.003	7462.3	12.9	***
Block 2	-0.009	0.003	7460.4	-2.85	**
Probe word x Validity	-0.011	0.002	7380.2	-4.82	***
Probe word x Block 1	-0.010	0.003	7461.4	-3.29	**
Probe word x Block 2	-0.001	0.003	7447.4	-0.27	
Validity x Block 1	0.004	0.003	7457.3	1.25	
Validity x Block 2	0.000	0.003	7461.8	0.10	
Probe word x Validity x Block 1	-0.006	0.003	7465.9	-1.98	*
Probe word x Validity x Block 2	-0.002	0.003	7463.8	-0.76	

Note. Validity (contrast coded: invalid = 1, valid = -1). Probe word (contrast coded: false = 1, correct = -1). Block 1 (contrast coded: Block 1 = 1, Block 2 = 0, Block 3 = -1). Block 2 (contrast coded: Block 1 = 0, Block 2 = 1, Block 3 = -1).

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

Table 2:

Estimated Coefficients, Standard Errors, and z-values for the Generalized Mixed Model of the Error Rates in Experiment 1.

	Est.	SE	z
(Intercept)	-4.472	0.172	-25.93 ***
Probe word	-0.224	0.091	-2.47 *
Validity	0.091	0.091	1.01
Block 1	-0.095	0.133	-0.72
Block 2	0.041	0.127	0.32
Probe word x Validity	-0.16	0.091	-1.77 .
Probe word x Block 1	-0.138	0.133	-1.03
Probe word x Block 2	0.071	0.127	0.56
Validity x Block 1	-0.198	0.133	-1.49
Validity x Block 2	0.136	0.127	1.07
Probe word x Validity x Block 1	-0.163	0.133	-1.23
Probe word x Validity x Block 2	0.097	0.127	0.76

Note. Validity (contrast coded: invalid = 1, valid = -1). Probe word (contrast coded: false = 1, correct = -1). Block 1 (contrast coded: Block 1 = 1, Block 2 = 0, Block 3 = -1). Block 2 (contrast coded: Block 1 = 0, Block 2 = 1, Block 3 = -1).

. $p < .1$, * $p < .05$, *** $p < .001$.

Table 3:

Estimated Coefficients, Standard Errors, Degrees of Freedom and t-values for the Linear Mixed Model of the Log-transformed Response Latencies in Experiment 2.

	Est.	SE	df	t	
(Intercept)	6.664	0.019	195.6	347.6	***
Probe word	0.018	0.003	17082.1	5.73	***
Validity	0.012	0.003	17058.3	3.78	***
Block 1	0.024	0.005	17165.5	5.24	***
Block 2	-0.015	0.004	17114.2	-3.33	***
Probe word x Validity	-0.036	0.003	17072.0	-11.35	***
Probe word x Block 1	-0.004	0.004	17066.2	-0.82	
Probe word x Block 2	-0.007	0.004	17069.0	-1.59	
Validity x Block 1	0.002	0.004	17085.0	0.55	
Validity x Block 2	-0.002	0.004	17071.0	-0.46	
Probe word x Validity x Block 1	-0.024	0.004	17083.3	-5.48	***
Probe word x Validity x Block 2	0.004	0.004	17067.9	0.91	

Note. Validity (contrast coded: invalid = 1, valid = -1). Probe word (contrast coded: false = 1, correct = -1). Block 1 (contrast coded: Block 1 = 1, Block 2 = 0, Block 3 = -1). Block 2 (contrast coded: Block 1 = 0, Block 2 = 1, Block 3 = -1).

*** $p < .001$.

Table 4:

Estimated Coefficients, Standard Errors, and z-values for the Generalized Mixed Model of the Error Rates in Experiment 2.

	Est.	SE	z	
(Intercept)	-3.089	0.11	-28.16	***
Probe word	0.069	0.029	2.38	*
Validity	0.041	0.029	1.42	
Block 1	0.287	0.040	7.16	***
Block 2	0.099	0.04	2.50	*
Probe word x Validity	-0.643	0.029	-21.86	***
Probe word x Block 1	-0.022	0.04	-0.56	
Probe word x Block 2	-0.041	0.039	-1.04	
Validity x Block 1	0.038	0.04	0.97	
Validity x Block 2	-0.004	0.039	-0.11	
Probe word x Validity x Block 1	-0.376	0.040	-9.48	***
Probe word x Validity x Block 2	0.020	0.039	0.51	

Note. Validity (contrast coded: invalid = 1, valid = -1). Probe word (contrast coded: false = 1, correct = -1). Block 1 (contrast coded: Block 1 = 1, Block 2 = 0, Block 3 = -1). Block 2 (contrast coded: Block 1 = 0, Block 2 = 1, Block 3 = -1).

* $p < .05$, *** $p < .001$.













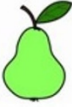



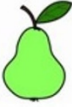



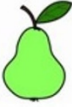



(a)	<table border="1"> <tr> <td>Picture</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Assertion</td> <td>This is a car</td> <td>This is a bike</td> <td>This is a bike</td> <td>This is a car</td> </tr> <tr> <td>Validity</td> <td>Valid</td> <td>Valid</td> <td>Invalid</td> <td>Invalid</td> </tr> </table>				Picture					Assertion	This is a car	This is a bike	This is a bike	This is a car	Validity	Valid	Valid	Invalid	Invalid
Picture																			
Assertion	This is a car	This is a bike	This is a bike	This is a car															
Validity	Valid	Valid	Invalid	Invalid															
(b)	<table border="1"> <tr> <td>Picture</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Assertion</td> <td>This is a pear</td> <td>This is a pineapple</td> <td>This is a pineapple</td> <td>This is a pear</td> </tr> <tr> <td>Validity</td> <td>Valid</td> <td>Valid</td> <td>Invalid</td> <td>Invalid</td> </tr> </table>				Picture					Assertion	This is a pear	This is a pineapple	This is a pineapple	This is a pear	Validity	Valid	Valid	Invalid	Invalid
Picture																			
Assertion	This is a pear	This is a pineapple	This is a pineapple	This is a pear															
Validity	Valid	Valid	Invalid	Invalid															

Figure 1. Examples of experimental items in four different versions (four combinations of pictures with assertions; two valid, two invalid) from the categories of vehicles (a) and fruit (b).

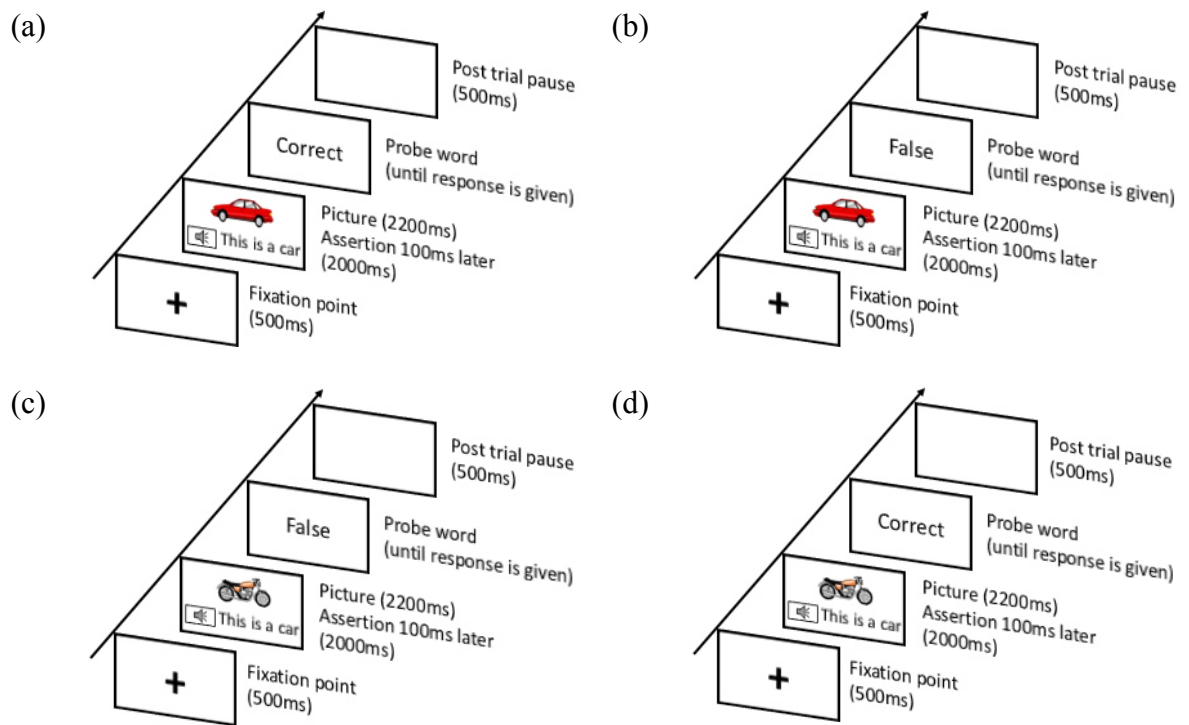
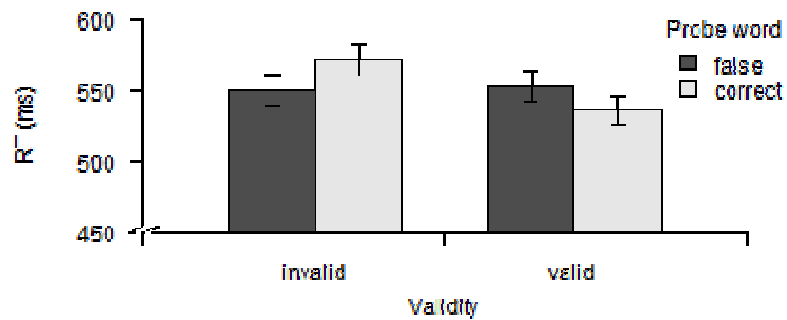
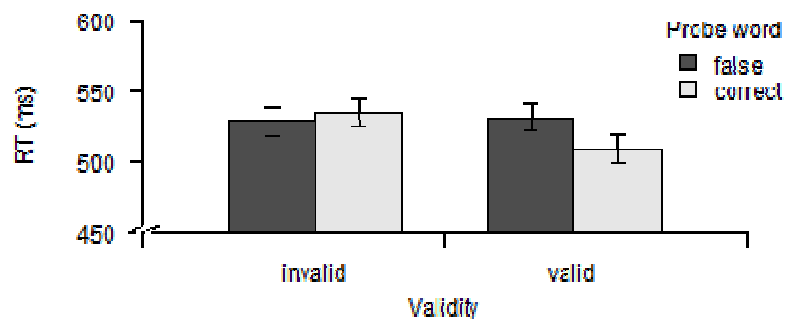


Figure 2. Structure of (a) a congruent trial with a valid picture-assertion combination and the probe word “correct”, (b) an incongruent trial with a valid picture-assertion combination and the Probe word “false”, (c) a congruent trial with an invalid picture-assertion combination and the probe word “false”, and (d) an incongruent trial with an invalid picture-assertion combination and the probe word “correct”.

(a)



(b)



(c)

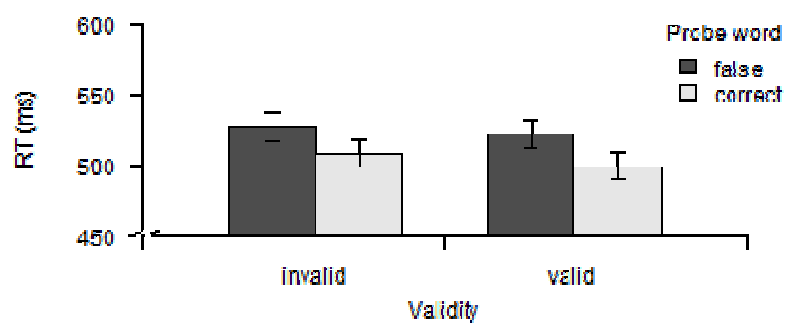
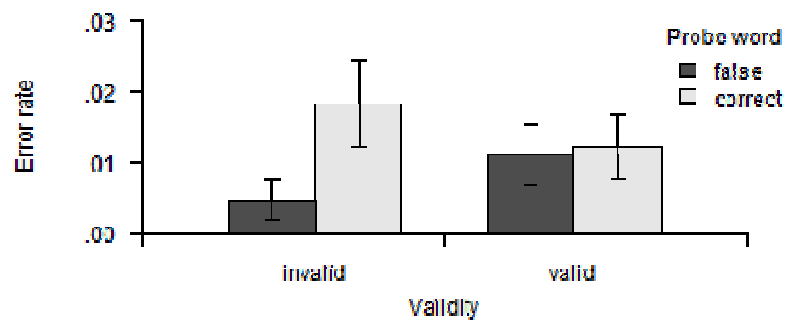
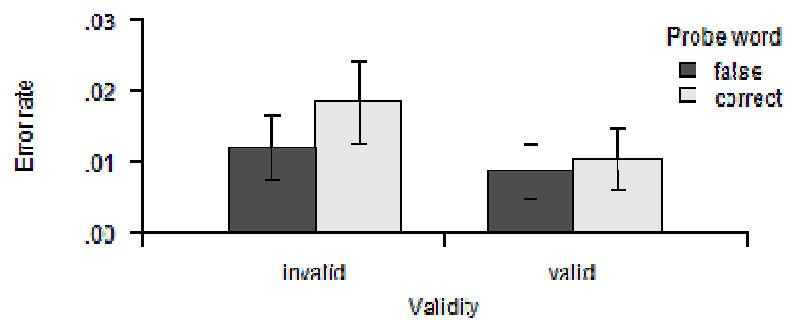


Figure 3. Back-transformed response latency estimated by the linear mixed model as a function of validity of the picture-assertion combination (invalid vs. valid) and probe word (false vs. correct) for (a) Block 1, (b) Block 2, and (c) Block 3 in Experiment 1 (adults). Error bars correspond to ± 1 standard error.

(a)



(b)



(c)

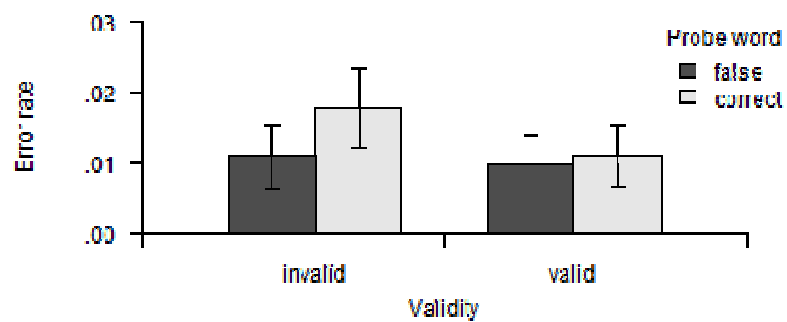
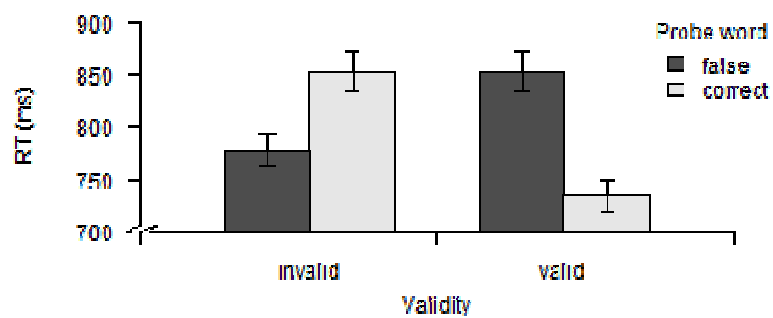
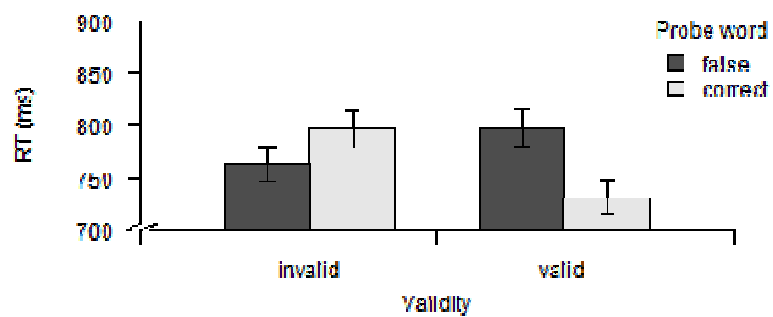


Figure 4. Back-transformed error probability estimated by the generalized linear mixed model as a function of validity of the picture-assertion combination (invalid vs. valid) and probe word (false vs. correct) for (a) Block 1, (b) Block 2, and (c) Block 3 in Experiment 1 (adults). Error bars correspond to ± 1 standard error.

(a)



(b)



(c)

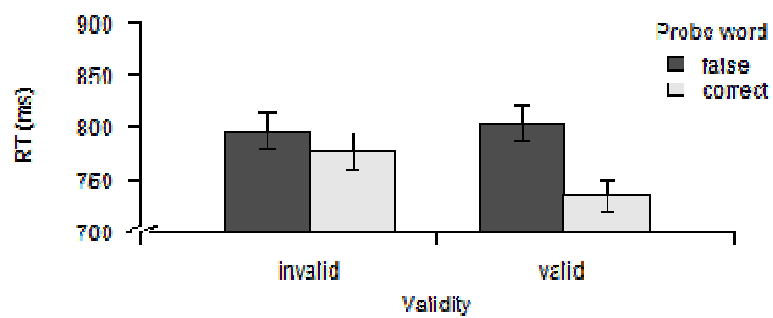
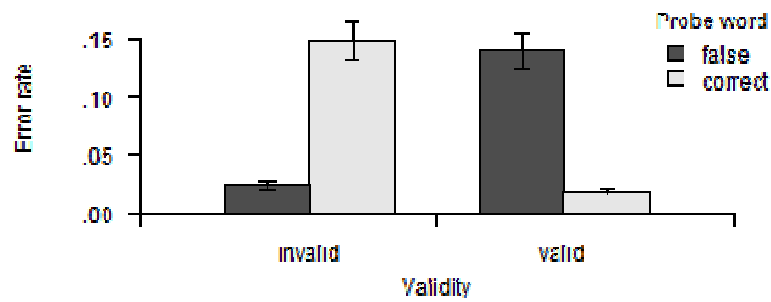
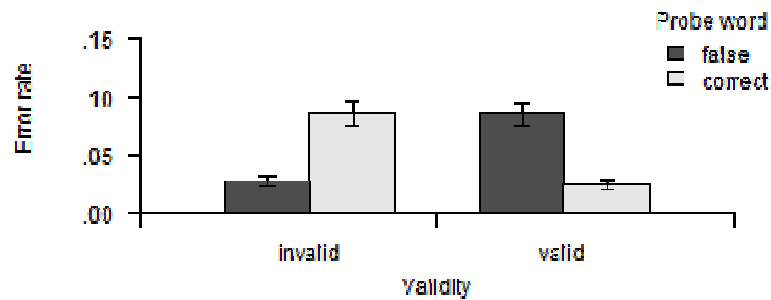


Figure 5. Back-transformed response latency estimated by the linear mixed model as a function of validity of the picture-assertion combination (invalid vs. valid) and probe word (false vs. correct) for (a) Block 1, (b) Block 2, and (c) Block 3 in Experiment 2 (children). Error bars correspond to ± 1 standard error.

(a)



(b)



(c)

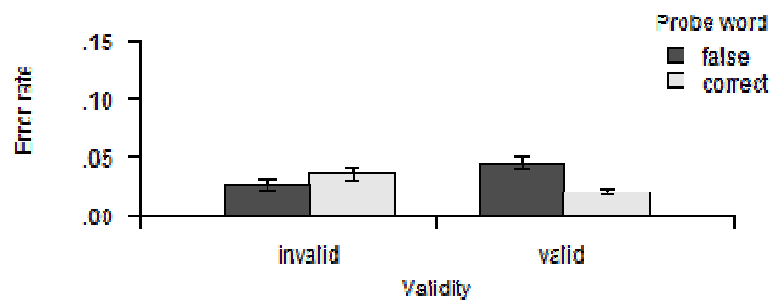


Figure 6. Back-transformed error probability estimated by the generalized linear mixed model as a function of validity of the picture-assertion combination (invalid vs. valid) and probe word (false vs. correct) for (a) Block 1, (b) Block 2, and (c) Block 3 in Experiment 1 (adults). Error bars correspond to ± 1 standard error.